N° d'ordre NNT : 2022LYSE2034

THÈSE de DOCTORAT DE L'UNIVERSITÉ DE LYON

Opérée au sein de

L'UNIVERSITÉ LUMIÈRE LYON 2

École Doctorale : ED 512 Informatique et Mathématiques

Discipline : Informatique

Soutenue publiquement le 7 juillet 2022, par :

Abderrazek AZRI

Approches multimodales d'apprentissage automatique pour la détection des rumeurs dans les *microblogs*.

Devant le jury composé de :

Christophe GRAVIER, Professeur des universités, Université Jean Monnet, Président

Pascale KUNTZ-COSPEREC, Professeure des universités, Université de Nantes, Rapporteur

Cédric WEMMERT, Professeur d'université, Université de Strasbourg, Rapporteur

Guillaume CABANAC, Maître de conférences HDR, Université Paul Sabatier - Toulouse 3, Examinateur

Nouria HARBI, Maîtresse de conférences, Université Lumière Lyon 2, Examinatrice

Cécile FAVRE, Maîtresse de conférences, Université Lumière Lyon 2, Examinatrice

Jérôme DARMONT, Professeur des universités, Université Lumière Lyon 2, Directeur de thèse

Contrat de diffusion

Thèse présentée pour obtenir le grade de
Docteur de l'Université Lumière Lyon 2

École Doctorale Informatique et Mathématiques (ED 512)
Unité de Recherche ERIC
Discipline : Informatique

Approches multimodales d'apprentissage automatique pour la détection des rumeurs dans les *microblogs*

Par : Abderrazek Azri

Présentée et soutenue publiquement le 07 juillet 2022, devant le jury composé de :

Pascale Kuntz, Professeure des Universités, Univ. Nantes	Rapportrice
Cédric Wemmert, Professeur des Universités, Univ. de Strasbourg	Rapporteur
Guillaume Cabanac, Maître de Conférences HDR, Univ. Toulouse 3	Examinateur
Christophe Gravier, Professeur des Universités, Univ. de St-Étienne	Examinateur
Cécile Favre, Maîtresse de Conférences, Univ. Lyon 2	Co-directrice
Nouria Harbi, Maîtresse de Conférences, Univ. Lyon 2	Co-directrice
Jérôme Darmont, Professeur des Universités, Univ. Lyon 2	Directeur

Résumé

La nature ubiquitaire et la popularité des plateformes de médias sociaux entraînent la génération d'une grande quantité de données multimédia. L'ouverture et la facilité de partage sur ces plateformes favorisent la diffusion de l'information sans nécessairement tenir compte de sa véracité. Sans moyen de vérification et d'analyse du contenu échangé, les rumeurs peuvent affecter sérieusement la crédibilité de ces plateformes et engendrer des conséquences désastreuses dans la vie réelle.

L'objectif de cette thèse est de proposer des solutions au problème de la classification automatique de la véracité des rumeurs dans les sites de microblogging. En se basant sur le contenu riche et de nature diverse fourni par ces plateformes nous considérons plusieurs caractéristiques des messages, notamment le texte, le contexte social, les sentiments et les images pour analyser leur véracité.

Nous proposons un cadre pour la fusion multimodale de caractéristiques qui s'appuie sur plusieurs modèles d'apprentissage automatique pour l'évaluation de la véracité des messages, ainsi que son implémentation dans le framework MONITOR (Multimodal Fusion Framework to Assess Message Veracity in Social Networks). L'originalité de ce travail réside notamment dans le recours à des indicateurs de qualité d'image comme caractéristiques des images pour qu'ils soient utilisés dans un contexte de détection de rumeurs.

Afin d'améliorer les performances de MONITOR, nous exploitons également l'apprentissage ensembliste, un paradigme très peu exploré dans le domaine de la classification des rumeurs. Nous proposons plusieurs algorithmes de meta-learning en utilisant les modèles individuels de MONITOR comme modèles de base.

Toujours avec l'idée d'améliorer la performance de la classification des rumeurs, nous proposons un modèle multimodal appelé deepMONITOR, basé sur les réseaux de neurones profonds, qui se révèlent capables d'apprendre efficacement des représentations de textes et d'images. L'apport de ce travail concerne particulièrement l'intégration des images et de l'analyse de sentiments dans l'apprentissage multimodal.

Enfin, pour répondre au problème du peu de jeux de données disponibles pour la détection multimodale des rumeurs, nous avons construit DAT@Z21, un jeu de données multimédia volumineux issues de Twitter, avec un étiquetage qui s'appuie sur une vérité terrain collectée à partir d'un site de fact checking. Ce jeu de données est partagé en respectant la réglementation en vigueur.

Les expériences que nous avons menées à l'aide de divers jeux de données, dont DAT@Z21, démontrent la pertinence de nos propositions. En effet, MONITOR et deepMONITOR réalisent les meilleures performances en les comparant avec les méthodes récentes de l'état de l'art.

Mots clés : Rumeurs, Microblog, Apprentissage multimodal, Apprentissage supervisé, Apprentissage ensembliste, Apprentissage profond, Ingénierie des caractéristiques, Métriques de qualité d'image, Analyse de sentiments, Jeux de données.

Abstract

The ubiquitous nature and popularity of social media platforms results in the generation of a large amount of multimedia data. The openness and ease of sharing on these platforms promotes the dissemination of information without necessarily considering its veracity. Without any means of verification and analysis of the content exchanged, rumors can seriously affect the credibility of these platforms and lead to disastrous consequences in real life.

The objective of this thesis is to propose solutions to the problem of automatic classification of the veracity of rumors in microblogging sites. Based on the rich and diverse content provided by these platforms, we consider several features of the messages, including text, social context, sentiment and images to analyze their veracity.

We propose a framework for multimodal feature fusion that relies on several machine learning models for message veracity assessment, as well as its implementation in the MONITOR (Multimodal Fusion Framework to Assess Message Veracity in Social Networks) framework. The originality of this work lies in the use of image quality indicators as characteristics of images to be used in a rumor detection context.

To improve the performance of MONITOR, we also exploit the ensemble learning, a paradigm that has been little explored in the field of rumor classification. We propose several meta-learning algorithms using MONITOR's individual models as base models.

Still with the idea of improving the performance of rumor classification, we propose a multimodal model called deepMONITOR, based on deep neural networks, which are shown to be capable of efficiently learning text and image representations. The contribution of this work concerns particularly the integration of images and sentiment analysis in multimodal learning

Finally, to address the problem of scarce datasets available for multimodal rumor detection, we built DAT@Z21, a large multimedia dataset from the Twitter microblog, with labels based on ground truth collected from a fact checking site. This dataset is shared in compliance with current regulations.

The experiments we conducted using various datasets, including DAT@Z21, demonstrate the relevance of our proposals. Indeed, MONITOR and deepMONITOR achieve the best performances compared to recent state-of-the-art methods.

Keywords : Rumors, Microblog, Multimodal learning, Supervised learning, Ensemble learning, Deep learning, Feature engineering, Image quality metrics, Sentiment analysis, Datasets.

Dédicaces

A mes chers parents,
A ma chère femme, qui m'a soutenu dans
les moments les plus difficiles de cette thèse,
A mes chers enfants : Oussama, Ayoub, Abderrahamane et Arwa,
A mes chers sœurs et frères,
A toute ma famille,
A tous mes amis.

Remerciements

Je tiens à exprimer toute ma reconnaissance et gratitude à mes encadrants, Cécile Favre, Nouria Harbi et Jérôme Darmont pour m'avoir accueilli au sein du laboratoire ERIC où j'ai pu réaliser cette thèse dans de bonnes conditions, et pour la confiance et la grande liberté intellectuelle et d'action qu'ils m'ont accordé pour mener à bien mes réflexions. Je souhaite, encore, les remercier pour leur aide et conseils avisés tout au long de cette thèse, leur gentillesse et leur constante disponibilité.

J'adresse tous mes remerciements aux membres du jury, Madame Pascale Kuntz, Professeure des Universités, Université de Nantes, ainsi qu'à Monsieur Cédric Wemmert, Professeur des Universités, Université de Strasbourg, de l'honneur qu'ils m'ont fait en acceptant d'être rapporteurs de cette thèse.

J'exprime également mes remerciements à Monsieur Christophe Gravier, Professeur des Universités, Université de St-Étienne pour avoir accepté d'évaluer ces travaux en tant qu'examinateur. Tous mes remerciements également à Monsieur Guillaume Cabanac, Maître de Conférences HDR, Université Toulouse 3, d'avoir accepté d'examiner ces travaux, mais d'avoir également pu faire part de retours très pertinents et bénéfiques à l'occasion des comités de suivi de thèse.

Je tiens à exprimer mes remerciements et ma reconnaissance à Monsieur Mohammed Salah Benbicha, pour son aide précieuse et ses encouragements à poursuivre mes études par la réalisation d'une thèse de doctorat.

Cette thèse n'aurait pu avoir lieu sans le financement du Ministère de la Défense Algérien que je remercie chaleureusement.

Je tiens à remercier tous les membres du laboratoire ERIC, permanents, doctorants, stagiaires (en particulier les membres de l'équipe SID) et la gestionnaire Habiba Osman, pour l'ambiance de travail très agréable pendant ces quatre années passées en votre compagnie. Un remerciement spécial à Ricco Rakotomalala pour ses précieux conseils.

Table des matières

Liste des figures

Liste des tableaux

Chapitre 1

Introduction

Sommaire

1.1 Préambule

Le développement explosif du Web depuis le milieu des années 1990 a fait évoluer considérablement la façon dont les individus communiquent entre eux. Les médias sociaux en ligne, comme Twitter [1] et Facebook [2], peuvent facilement diffuser des informations en temps réel provenant d'utilisateurs du monde entier. Avec leurs avantages en termes de facilité d'utilisation, de faible coût et de rapidité de diffusion, les médias sociaux sont devenus une scène ouverte pour la discussion, l'expression d'idéologies, la diffusion de connaissances, le partage d'émotions et de sentiments en ligne. En effet, selon le *Global Digital Report 2021* [3], sur une population mondiale totale de 7,8 milliards d'habitants, on compte 4,7 milliards d'utilisateurs d'internet et 4,2 milliards d'utilisateurs actifs de médias sociaux.

À cause de leur popularité croissante, les médias sociaux en ligne deviennent un terrain idéal pour la diffusion des fausses nouvelles (*fausses informations*). Ces informations trompeuses sont désormais plus répandues et largement diffusées sur les médias sociaux que dans les médias traditionnels. Étant souvent utilisées pour semer la confusion et persuader les utilisateurs en ligne sur la base de faits biaisés, les fausses nouvelles peuvent constituer une menace pour la crédibilité d'Internet et entraîner de graves conséquences dans la vie réelle.

La portée des fausses informations a été particulièrement mise en évidence pendant la campagne des élections présidentielles américaines de 2016. Les fausses nouvelles ont été accusées d'accroître la polarisation politique et les conflits partisans pendant la campagne électorale, et les électeurs ont été facilement influencés par les déclarations et affirmations politiques trompeuses. Lors de ces élections, le nombre de partages, réactions et commentaires générés suite au partage et la diffusion de fausses informations ont dépassé sur Facebook seul celui généré sur l'ensemble de dix-neuf grands sites d'information.

Face à ces constatations, nous sommes donc amenés à formuler les questions suivantes. Jusqu'à quel point et dans quelle mesure les faits diffusés sont vérifiés ? À quel point pouvons-nous nous fier au contenu de l'information que nous consultons tous les jours ? La question des fausses informations en ligne est devenue une préoccupation majeure

1. https://twitter.com/
2. https://www.facebook.com/
3. https://datareportal.com/reports/digital-2021-october-global-statshot

récemment, de la part des praticiens (journalistes et professionnels de l'information et de la communication) et du monde académique. De grands efforts ont été déployés pour lutter contre la diffusion des fausses informations.

Le reste de ce chapitre est organisé comme suit. Dans la section 1.2, nous apportons une définition aux termes « fausse information ». Dans la section 1.3, nous discutons les facteurs techniques et les fondements psychologiques et cognitifs favorisant la prolifération des fausses nouvelles. Nous présentons les motivations ainsi que la problématique de cette thèse dans la section 1.4. Enfin, dans la section 1.5, nous présentons les contributions et l'organisation de ce manuscrit de thèse.

1.2 Définition des termes « fausse information » et « rumeur »

Afin de cerner l'objet de notre recherche, il apparaît nécessaire d'apporter préalablement une définition des termes « fausse information » ou « fausse nouvelle ».

Les fausses informations[4] existent depuis très longtemps, pratiquement depuis que les nouvelles ont commencé à circuler largement après l'invention de la presse écrite. Cependant, il n'existe pas de définition universelle de ces termes, même dans le milieu journalistique. Pour éviter toute ambiguïté, nous commençons d'abord par (1) présenter une définition et (2) distinguer théoriquement plusieurs concepts qui sont fréquemment associés aux fausses informations ou qui se croisent avec elles, notamment le terme rumeur.

Bien qu'initialement utilisés par certaines personnes pour faire référence à une information dépourvue de source ou un avis qui ne peut être vérifié, ces termes ont évolué et sont devenus synonymes de diffusion de fausses informations [Cooke, 2017]. Une définition adoptée par la majorité des travaux de la littérature [Shu et al., 2017; Allcott and Gentzkow, 2017] est la suivante.

Définition 1 (Fausses informations) *Articles de nouvelles créés intentionnellement, vérifiables comme étant faux et qui pourraient induire les lecteurs en erreur.*

4. Définition du mot *news* dans le dictionnaire Oxford : informations récentes ou dignes d'intérêt, notamment sur des événements récents.

Cette définition comporte deux caractéristiques essentielles, la véracité et l'intention. Premièrement, les *fake news* comprennent des informations fausses qui peuvent être vérifiées comme telles. Deuxièmement, les fausses informations sont créées avec l'intention malhonnête de tromper leurs lecteurs.

Les études existantes utilisent souvent et de manière interchangeable les termes fausses informations pour désigner divers concepts tels que mésinformation, désinformation, information satirique et rumeur. Sur la base de la définition de ces termes et concepts, nous pouvons les distinguer les uns des autres en fonction des deux caractéristiques présentées dans la définition 1 : (1) la véracité (fausse ou non) et (2) l'intention (mauvaise ou bonne). Par exemple, la désinformation est une information fausse et mal intentionnée qui vise à tromper les utilisateurs. Le tableau 1.1 illustre la comparaison entre ces concepts.

TABLE 1.1 – Comparaison des concepts liés aux termes « fausses informations »

Concepts	**Véracité**	**Intention**
Désinfomation	Fausse	Mauvaise
Mésinfomation	Fausse	Inconnue
Information satirique	Inconnue	Bonne
Rumeur	Inconnue	Inconnue

Par ailleurs, de nombreuses définitions différentes du terme rumeur ont été proposées dans la littérature. [Zhao et al., 2015] définissent une rumeur comme étant une déclaration controversée et vérifiable, Zubiaga et al. [2018] comme une information en circulation dont le statut de véracité n'a pas encore été vérifié au moment de la publication. Pour [Hamidian and Diab, 2015], une rumeur peut être à la fois vraie et fausse. C'est une affirmation dont la véracité est mise en doute et n'a pas de source claire, même si ses origines et ses intentions idéologiques ou partisanes sont claires.

[Cao et al., 2018] vont plus loin en proposant des familles de rumeurs. Les rumeurs générales ont valeur de vérité non vérifiée. Elles englobent deux sous-familles de rumeurs : les rumeurs objectives, dont la valeur de vérité est confirmée par une source fiable ou crédible, et les rumeurs subjectives, dont la valeur de vérité est déterminée par les jugements subjectifs des utilisateurs des réseaux sociaux.

Ces définitions partagent généralement deux idées communes sur la nature des rumeurs. Primo, une des caractéristiques des rumeurs est qu'elles apparaissent dans un contexte d'ambiguïté. Par conséquent, leur véracité est incertaine. Secundo, bien que

sa valeur de véracité soit incertaine, une rumeur n'implique pas forcément de fausses informations.

1.3 Origine de la diffusion de rumeurs

De nos jours, la diffusion des nouvelles s'est déplacée vers les médias sociaux en ligne, où il est plus facile de consulter, partager et discuter des actualités avec des amis ou d'autres utilisateurs. Comme il est plus facile et plus rapide de produire du contenu en ligne, les barrières pour accéder aux médias en ligne ont considérablement diminué. Cela a entraîné la diffusion d'informations de faible qualité, qui ne respectent pas les normes journalistiques traditionnelles. Ces informations ne font pas l'objet d'un filtrage et d'une vérification des faits par des tiers. Ces facteurs, ainsi que le déclin de la confiance générale dans les médias de masse traditionnels, sont les principaux moteurs de la croissance explosive des fausses nouvelles sur les médias sociaux. Les deux principales motivations qui ont été avancées pour l'essor des sites de désinformation sont : (1) une motivation financière, les articles d'actualité viraux générant d'importants revenus publicitaires, et (2) une motivation plus idéologique, car les diffuseurs de fausses nouvelles cherchent généralement à influencer l'opinion publique sur des sujets particuliers [Allcott and Gentzkow, 2017].

Outre les aspects techniques des plateformes de réseaux sociaux favorisant la diffusion de fausses informations, un ensemble d'aspects psychologiques, cognitifs et sociaux sont aussi considérés comme des facteurs clés de la prolifération des fausses nouvelles. Les êtres humains ne sont pas formés pour faire la différence entre les vraies et les fausses nouvelles. Il existe plusieurs théories psychologiques et cognitives qui peuvent expliquer ce phénomène et le pouvoir d'influence des fausses nouvelles. Il y a deux facteurs majeurs qui rendent les utilisateurs naturellement vulnérables aux fausses nouvelles. (1) Le réalisme naïf : les utilisateurs ont tendance à croire que leurs perceptions de la réalité sont les seuls points de vue exacts, tandis que ceux qui ne sont pas d'accord sont considérés comme mal informés, irrationnels ou partiaux [Ross et al., 1996]. (2) Le biais de confirmation : les utilisateurs préfèrent recevoir des informations qui confirment leurs opinions [Nickerson, 1998]. En raison de ces biais cognitifs inhérents à la nature humaine, les fausses nouvelles peuvent souvent être perçues comme réelles par les lecteurs.

Certaines études mentionnent également l'importance de la théorie de l'identité sociale [Ashforth and Mael, 1989] et de l'influence sociale normative [Asch, 1951]. Ainsi,

les utilisateurs ont tendance à effectuer des actions qui sont socialement plus sûres, en consommant et en diffusant des informations qui sont en accord avec les normes établies au sein de leurs communautés, même si l'information partagée est fausse. Tous ces facteurs contribuent dans une certaine mesure à la création de l'effet de « chambre d'écho » qui donne lieu à la formation de groupes homogènes où les individus se ressemblent, partagent et discutent des mêmes idées. Ces groupes sont généralement caractérisés par des opinions extrêmement polarisées, car ils sont isolés des opinions et des perspectives contraires. Ces communautés étroitement liées sont le principal vecteur de diffusion des fausses informations [Del Vicario et al., 2016].

1.4 Motivations et problématique de la thèse

Le problème des fausses informations en ligne a suscité une attention croissante de la part des chercheurs et des professionnels de l'information et de la communication (journalistes). Premièrement, des approches basées sur la détection manuelle ont été proposées par les professionnels, qui se sont traduites, notamment, par l'apparition de plusieurs sites de vérification de faits (*fact checking*) ou de plateformes de la vérification participative (*crowdsourced fact checking*). Nous revenons avec plus de détails sur ces solutions pratiques proposées par les professionnels dans le chapitre suivant.

Malgré leur apport important dans la détection des fausses informations, ces approches souffrent, cependant, de deux difficultés majeures, particulièrement dans un contexte des réseaux sociaux en ligne :

— Avec une grande quantité d'informations générée par les réseaux sociaux, la vérification manuelle devient une tâche laborieuse et coûteuse.
— L'identification d'une nouvelle rumeur implique des compétences spécifiques et des analyses plus approfondies. Pratiquement, il n'est pas trivial pour un expert en journalisme de vérifier les faits d'un article d'actualité, sans parler des non-professionnels (le cas du *crowdsourcing*), ce qui pose le problème de la qualité de la vérification (le label).

Pour remédier à ces problèmes, de nombreuses recherches ont été menées ces dernières années afin de développer des solutions automatiques pour la détection et la classification des rumeurs dans les réseaux sociaux. Avec les avancées récentes réalisées dans le domaine de l'apprentissage automatique et l'apprentissage profond, des outils puissants qui sont

capables d'intégrer le langage naturel, de mémoriser les ordres séquentiels sémantiques importants et de capturer les relations sémantiques sous-jacentes, ont été largement utilisé dans le domaine de la classification automatique des rumeurs.

Par rapport à la vérification manuelle des faits, ces techniques automatiques sont en mesure de fournir de meilleurs résultats avec une grande efficacité, car ils n'ont pas besoin de labéliser toutes les données. Cependant, la détection automatique des rumeurs demeure une tâche très difficile, car les données des médias sociaux sont en grande partie non-structurées, incomplètes et bruitées, ce qui rend leur traitement et leur compréhension très difficiles. Nous résumons ci-dessous les problèmes auxquels sont confrontés les techniques de détection automatiques.

1. **Compréhension de la sémantique :** la plupart des rumeurs sont fabriquées intentionnellement pour tromper les utilisateurs. De fausses informations sont mélangées et déguisées dans une rumeur. Il est très difficile pour une machine de capter complètement cette sémantique.
2. **De grandes variations :** les rumeurs peuvent couvrir toutes sortes de sujets et prendre différents styles linguistiques. Les caractéristiques qui sont efficaces pour distinguer un certain type de rumeurs peuvent être peu performantes pour d'autres types. De plus, les algorithmes entraînés sur des données labélisées peu nombreuses peuvent échouer à détecter de nouvelles rumeurs.
3. **Données multimodales :** outre le texte pur, les rumeurs peuvent prendre la forme d'images ou de vidéos. Bien que des informations provenant de différentes modalités (type ou nature) puissent fournir des indices pour la détection des rumeurs, la question relative à la manière d'extraire des caractéristiques significatives de chaque modalité et de les fusionner efficacement constitue un véritable défi.
4. **Structure de propagation hétérogène :** pendant la diffusion des rumeurs sur un réseau social, les utilisateurs peuvent interagir et émettre des commentaires. Les comportements des utilisateurs aident à construire un contexte social sous-jacent aux rumeurs. La caractérisation de ces comportements sur le réseau de propagation est cruciale pour identifier les rumeurs en ligne.
5. **Besoin en données labélisées :** le manque de jeux de données complets et labélisés est une limitation majeure pour l'expérimentation et la validation de techniques supervisées. En effet, les jeux de données existants sont peu nombreux et aucun d'entre eux ne fournit toutes les caractéristiques qui se sont avérées efficaces pour l'analyse des rumeurs.

Motivés d'une part par le succès grandissant des techniques d'apprentissage automatique dans plusieurs domaines et, d'autre part, par les données riches en informations offertes par les sites des réseaux sociaux comme le texte, le contexte social de la diffusion, l'information visuelle, etc., nous avons décidé d'explorer plusieurs paradigmes d'apprentissage automatique et profond pour apprendre des informations issues de ces modalités. Nous réservons une attention particulière au contenu visuel, notamment les images, dont le potentiel pour la détection des rumeurs demeure insuffisamment exploité par les travaux de recherche. Les images sont largement répandues sur les sites de *microblogging*, présentent un contenu riche en informations et jouent un rôle important dans la diffusion des rumeurs.

Pour terminer, un *microblog* est défini comme « un site de média social sur lequel un utilisateur publie des messages courts et fréquents » selon le dictionnaire Oxford English. Les travaux présentés dans cette thèse utilisent des données issues du *microblog* Twitter car, ces dernières années, Twitter est devenu la source de données par excellence pour la collecte et l'analyse de rumeurs, et ce grâce à l'ouverture de son API par rapport aux restrictions imposées par Facebook et Sina Weibo [5].

1.5 Contributions de la thèse et organisation du manuscrit

Nos travaux de thèse ont abouti à plusieurs contributions. Premièrement, nous présentons et discutons dans le chapitre 2 l'état de l'art lié aux travaux présentés dans ce manuscrit. Cet état de l'art nous conduit à proposer une cartographie des solutions développées par le monde académique pour la prédiction de la véracité des rumeurs. En vue d'associer le contenu visuel des messages dans l'analyse de la véracité des rumeurs, nous présentons et discutons une typologie des approches de vérification de la véracité d'une image.

Deuxièmement, pour évaluer la véracité des messages postés sur les réseaux sociaux, nous proposons dans le chapitre 3 une plateforme de fusion multimodale appelée MONITOR, qui utilise des caractéristiques extraites du contenu textuel du message, du contexte social, ainsi que des caractéristiques des images. Pour représenter les images, nous utilisons un ensemble de caractéristiques avancées, inspirées du domaine

5. https://weibo.com/

de l'évaluation de la qualité des images. Plus précisément, toutes ces modalités sont fusionnées et plusieurs algorithmes d'apprentissage automatique sont utilisés pour la classification.

Troisièmement, nous proposons dans le chapitre 4 une extension à MONITOR en explorant un paradigme non-conventionnel d'apprentissage automatique pour la classification des rumeurs. Il s'agit de l'utilisation de plusieurs modèles de *metalearning* de l'apprentissage ensembliste. Ce type d'algorithmes a rarement été appliqué jusqu'ici pour la tâche de détection de rumeurs.

Quatrièmement, le chapitre 5 détaille deepMONITOR, un modèle de bout en bout qui repose sur des réseaux neuronaux profonds, en utilisant les trois caractéristiques des messages, en l'occurence le contenu textuel et visuel des messages, ainsi que les signaux sentimentaux qui sont très étroitement liés à l'analyse des rumeurs.

Cinquièmement, motivés par le besoin d'un jeu de données pour la classification des rumeurs dans les *microblogs*, nous construisons et partageons dans le chapitre 6 un jeu de données volumineux et complet appelé DAT@Z21, qui inclut toutes les caractéristiques nécessaires pour la classification de rumeurs, notamment des données textuelles et linguistiques, visuelles, spatio-temporelles, ainsi que des données relatives à l'engagement social et au comportement des utilisateurs.

Enfin, le chapitre 7 synthétise et discute les contributions effectuées durant la thèse. Nous y détaillons également les perspectives possibles à nos travaux.

Chapitre 2

État de l'art général

Sommaire

Publication(s) associée(s)

Azri, A., Favre, C., Harbi, N.,Darmont, J. : Vers une analyse des rumeurs dans les réseaux sociaux basée sur la véracité des images : état de l'art. 15ème journées EDA Business Intelligence & Big Data **EDA'2019**, volume B-15, Revue des Nouvelles Technologies de l'Information, pages 125-142, Montpellier, France, October 2019.

2.1 Introduction

Avec le développement immense d'Internet et des plateformes de réseaux sociaux, la diffusion massive de fausses nouvelles a le potentiel d'avoir des impacts extrêmement négatifs sur les individus et la société. Face à ces menaces, la détection des fausses nouvelles dans les médias sociaux est devenue ces dernières années un sujet de recherche actif attirant beaucoup d'attention de la part du monde du journalisme et de la communauté scientifique. Cet intérêt grandissant peut s'expliquer par les défis que cela représente, tant au niveau scientifique que sociétal.

Avant de présenter les approches dites pratiques proposées par les professionnels et les solutions automatiques existantes proposées par le monde académique, il est judicieux de comprendre certains aspects communs qui caractérisent les informations diffusées sur les réseaux sociaux en ligne. Malgré les spécificités très diverses de ces plateformes, les messages publiés par leurs utilisateurs partagent pratiquement toutes ces caractéristiques. La figure 2.1 illustre un exemple de message partagé par un utilisateur sur Twitter.

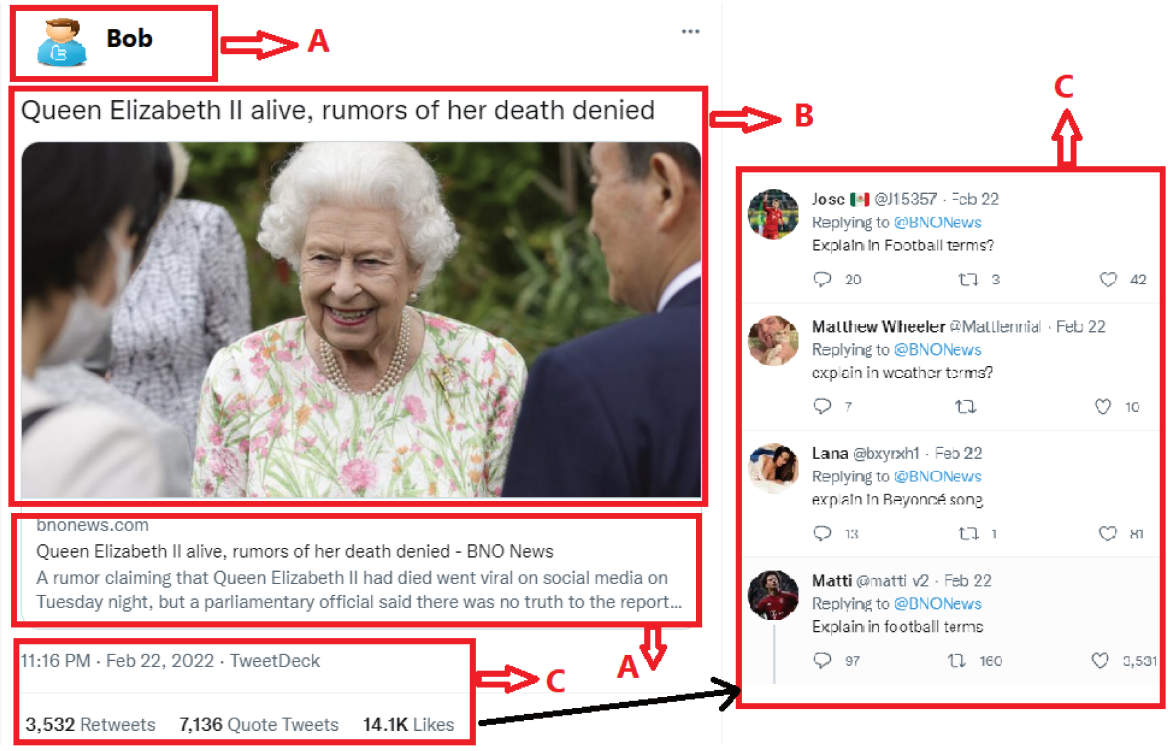


FIGURE 2.1 – Exemple de message Twitter : (A) L'utilisateur/diffuseur, (B) Le contenu, (C) Le contexte social

Chaque message peut être vu comme la combinaison des éléments suivants.

— **L'utilisateur/diffuseur** est la source de l'information (optionnelle) ou celui qui a procédé à la diffusion de l'information en ligne. Dans la partie D de la figure 2.1, nous pouvons voir que `bnonews.com` est la source de l'information et que l'utilisateur Bob est le diffuseur.
— **Le contenu** fait référence au corps de la publication (la partie B de la figure 2.1), il représente deux aspects.

1. **Un contenu textuel** décrivant l'événement d'actualité. Il peut fournir des détails sur l'événement et peut contenir certaines opinions ou certains sentiments à l'égard de cette nouvelle. Il peut contenir des liens vers des sources internet (URL), des émoticônes ou encore d'autres contenus informatifs liés au style de communication du média social, par exemple les *hashtags* (#) et les références vers des utilisateurs (@) sur Twitter.
2. **Un contenu visuel** qui peut correspondre, s'il existe, à une image ou à une vidéo. Il est souvent fourni pour compléter le contenu textuel et appuyer le sujet du message.

— **Le contexte social** (la partie C de la figure 2.1) reflète l'environnement social dans lequel évolue la diffusion de l'information, notamment les interactions entre les utilisateurs en ligne.

Sur la base de cette caractérisation des messages postés sur les réseaux sociaux, nous présentons une définition du problème de détection automatique des rumeurs. Formellement, une rumeur r est définie comme un ensemble $M = \{m_1, m_2, ..., m_n\}$ de n messages associés postés par un ensemble $U = \{u_1, u_2, ..., u_k\}$ de $k \leq n$ utilisateurs. Chaque message m_i est composé d'un n-uplet représentant le texte, les images, les vidéos et les autres contenus qu'il contient. Chaque utilisateur u_i est représenté par un ensemble d'attributs tels que le nom, l'âge de son compte, l'image avatar de son profil, etc. La tâche de détection des rumeurs est alors définie comme suit.

Définition 2 (Détection d'une rumeur) *Étant donné une rumeur r représentée par un ensemble de messages M et un ensemble d'utilisateurs U, la tâche de la classification de rumeur vise à déterminer si la rumeur peut être confirmée en tant que vraie ou fausse en apprenant une fonction de prédiction $\mathcal{F}(r) \longrightarrow \{0, 1\}$ telle que :*

$$\mathcal{F}(r) = \begin{cases} 1 & \text{si } r \textit{ est fausse,} \\ 0 & \text{sinon.} \end{cases}$$

Nous définissons donc la détection d'une rumeur comme un problème de classification binaire, pour deux raisons. (1) Cette définition s'accorde avec la majorité des travaux de la littérature. (2) Une rumeur est essentiellement un biais de distorsion sur l'information manipulée par l'utilisateur. Selon les recherches sur la théorie du biais médiatique [Gentzkow et al., 2015], le biais de distorsion est généralement modélisé comme un problème de classification binaire. Il y a lieu de noter que certains travaux ajoutent une troisième classe « non vérifiée » [Zubiaga et al., 2018] dans le cas où la véracité de la rumeur demeure indéterminée.

Contrairement à la définition 2, qui se concentre sur la détection de rumeur au niveau de l'événement associé, nous faisons partie des travaux qui se focalisent sur la détection de rumeur au niveau du message individuel. Plus précisément, étant donné un message qui prétend fournir des informations sur une rumeur, notre tâche consiste à prédire sa véracité comme étant fausse ou vraie en nous basant sur toutes les informations multimodales qu'il renferme.

Le reste de cet état de l'art est organisé comme suit. Dans la section 2.2, nous résumons les approches pratiques proposées par les journalistes. Nous présentons et discutons les différentes approches automatiques de détection des rumeurs dans les réseaux sociaux dans la section 2.3. Le contenu visuel des publications joue un rôle important dans la diffusion de l'information. À titre d'exemple, dans le cas de Twitter, les statistiques montrent qu'un *tweet* incluant une image obtient 150 % de *retweets*, 89 % de *likes* et 18 % de clics de plus qu'un *tweet* ne comportant aucune image[6]. Par conséquent, analyser la véracité du message implique nécessairement l'analyse de la véracité de l'image jointe. C'est ainsi que nous présentons dans la section 2.4 les techniques de détection de modifications dans une image numérique et que nous discutons leur applicabilité dans le contexte des réseaux sociaux. Enfin, nous clôturons ce chapitre sur une discussion des choix de méthodes pour les travaux présentés dans ce manuscrit.

6. https://www.blogdumoderateur.com/chiffres-twitter/

2.2 Approches pratiques de détection des fausses nouvelles

Dans cette section, nous résumons quelques approches pratiques déployées par les professionnels de l'information pour combattre les effets négatifs de la diffusion des fausses nouvelles. Une première étape consiste à analyser ou détecter manuellement les fausses nouvelles en utilisant le processus de vérification des faits (*fact checking*). Initialement développée dans le domaine du journalisme, elle vise à évaluer l'authenticité des informations en comparant les connaissances extraites du contenu de l'information à vérifier (par exemple, ses affirmations ou ses déclarations) avec des faits connus (c'est-à-dire des connaissances réelles). En indiquant aux utilisateurs ce qui est vrai, faux ou entre les deux, la vérification des faits est un bon moyen d'identifier les fausses nouvelles. De manière générale, la vérification manuelle des faits peut être divisée en deux catégories : (1) la vérification par des experts et (2) la vérification participative (*crowdsourced fact checking*).

La vérification des faits par des experts s'appuie sur des personnes (journalistes) du domaine (vérificateurs de faits). Elle est souvent effectuée par un petit groupe de vérificateurs hautement crédibles et donne des résultats très précis, mais elle est coûteuse et difficile à réaliser dans le cas d'un grand volume de contenus d'informations à vérifier. Récemment, de nombreux sites web sont apparus pour permettre la vérification des faits par des experts afin de mieux servir les usagers. Une liste exhaustive de sites web de vérification des faits est fournie par le *Reporters Lab* de l'Université de Duke[7], qui répertorie plus de deux cents sites web de vérification des faits dans différents pays et en différentes langues, dont plus d'une vingtaine en France (AFP's Factuel, CheckNews de Libération, L'instant détox de Franceinfo, Les Décodeurs du journal Le Monde, 20 Minutes Fake Off, etc.).

En général, ces sites de vérification des faits par des experts produisent une vérité de terrain sur l'authenticité des affirmations, des déclarations, des articles ou des messages, mais sans aucune information sur l'intention. De plus, il est très difficile, voire impossible, de quantifier ou détecter automatiquement l'intention d'une personne à diffuser volontairement ou non une fausse information. Le tableau 2.1 présente les détails de quelques sites web de vérification des faits.

7. https://reporterslab.org/fact-checking/#

TABLE 2.1 – Comparaison des sites web de vérification des faits par des experts

Site Web	**Sujets traités**	**Labels d'évaluation**	**API**
Politifact.com	Divers	Vrai, majoritairement vrai, à moitié vrai, majoritairement faux, faux, ridicule	Oui
Snopes.com	Politique et autres questions sociales et d'actualité	Vrai, majoritairement vrai, mélange, majoritairement faux, faux, non prouvé et autres labels	Non
TruthOrFiction.com	Divers	Vérité, non prouvée, vérité et fiction, vérité antérieure, contestée et autres labels	Non
FullFact.com	Divers	Ambiguïté (pas de labels clairs)	Non
FactCheck.org	Politique américaine	Vrai, pas de preuve, faux	Non
Hoax-Slayer.com	Divers	Canulars, arnaques, fausses alertes, fausses nouvelles, tromperie, vérité, humour, spams	Non

Certains sites web fournissent des informations supplémentaires, par exemple, PolitiFact fournit une carte de scores qui présente des statistiques sur l'authenticité de toutes les déclarations liées à un sujet spécifique (par exemple, Joe Biden, 46e Président des États-Unis – figure 2.2). Ces informations peuvent aider à identifier les sujets importants, qui nécessitent un examen plus approfondi pour la vérification. Le site web de PolitiFact est l'un des rares sites de vérification de faits qui fournit également une API (*Application Programming Interface*) pour que les utilisateurs puissent accéder au texte intégral des déclarations et des articles qui ont été vérifiés.

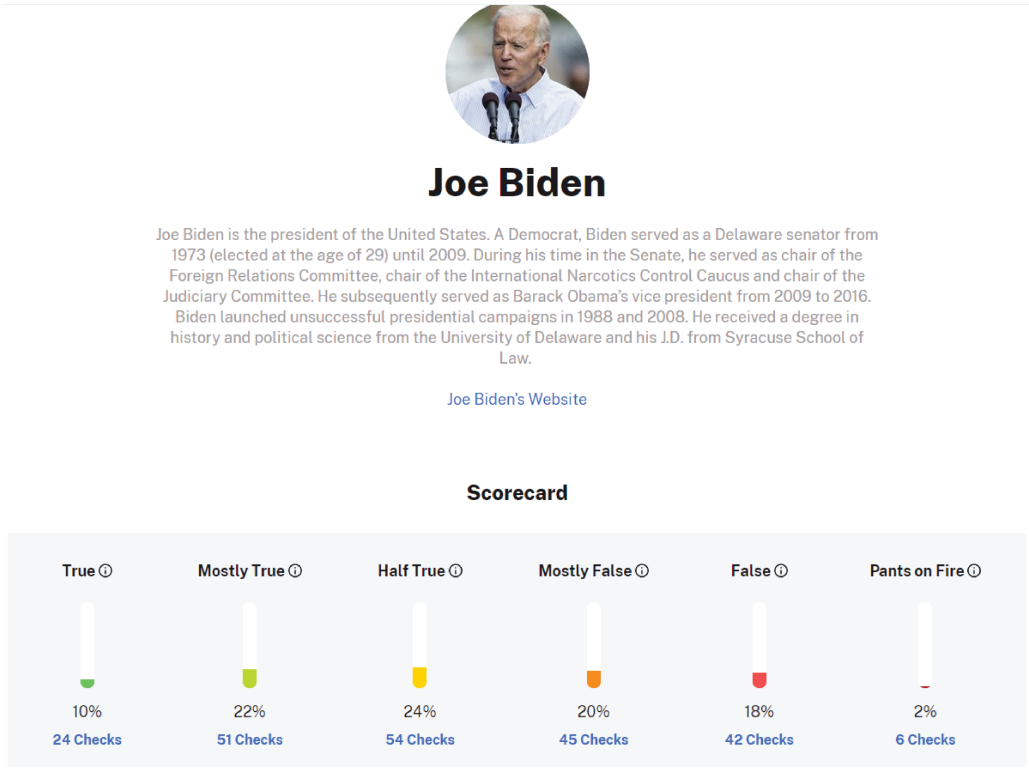


FIGURE 2.2 – Carte de scores de PolitiFact (vérification des faits par des experts)

Les analyses détaillées fournies par les experts de ces sites web suite à l'évaluation du contenu de l'information (qu'est-ce qui est faux et pourquoi est-ce faux ?) apportent des informations précieuses sur divers aspects de l'analyse des fausses nouvelles.

La vérification collaborative des faits, elle, s'appuie sur une large population d'individus agissant comme vérificateurs de faits. Par rapport à la vérification des faits par des experts, la vérification collective des faits est relativement difficile à gérer, moins crédible et moins précise en raison du biais politique des vérificateurs et de leurs annotations contradictoires, mais présente une meilleure scalabilité. Ainsi, dans la vérification collective des faits, il faut souvent : (1) filtrer les utilisateurs non crédibles et (2) résoudre les résultats contradictoires de la vérification. Néanmoins, les plateformes

de *crowdsourcing* permettent souvent de fournir aux vérificateurs des commentaires plus détaillés (par exemple, leurs sentiments ou leurs positions), qui peuvent être explorés après dans les études sur les fausses nouvelles.

Contrairement à la vérification des faits par des experts, les sites web de vérification collaborative des faits sont encore en phase de développement. Un exemple est Fiskkit [8], où les utilisateurs peuvent télécharger volontairement des articles, donner des notes aux phrases des articles et choisir les labels qui les décrivent le mieux. Les sources d'articles données aident à distinguer les types de contenu et leur crédibilité. Les labels catégorisés en plusieurs dimensions permettent d'étudier les tendances entre les articles d'information vrai ou faux (figure 2.3).

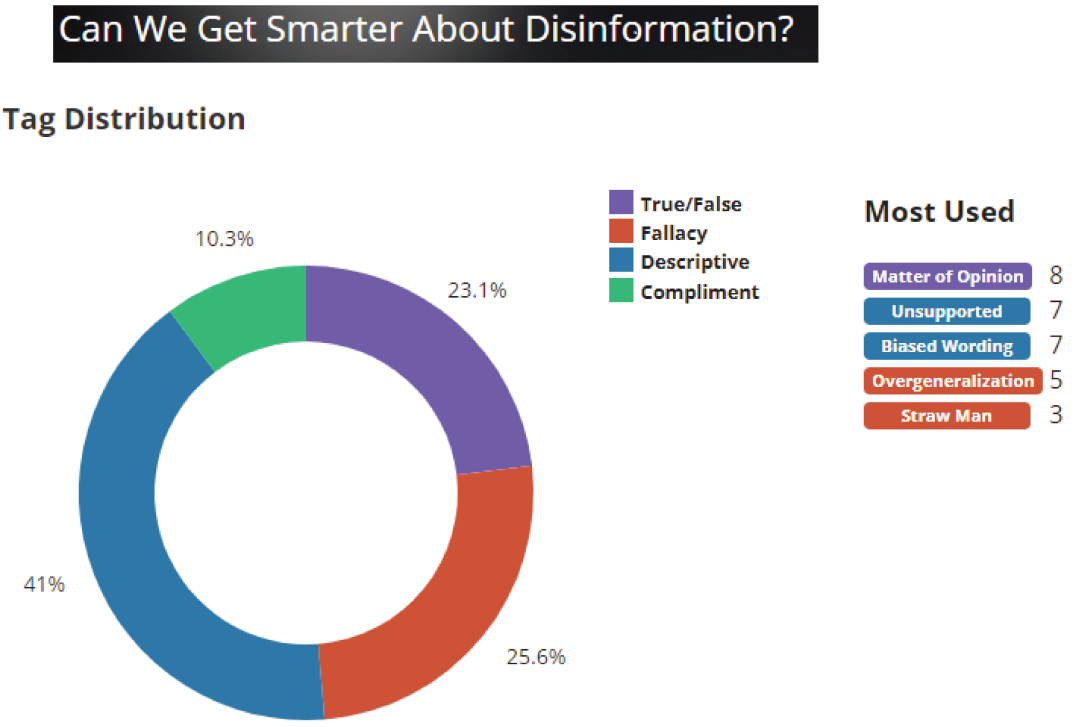


FIGURE 2.3 – Distribution des labels de Fiskkit (vérification collective des faits)

Malgré l'utilité de la vérification manuelle des faits, les ressources actuelles présentent plusieurs limites. Le processus de détection nécessite beaucoup de temps et un travail manuel important. Par conséquent, le temps que les fausses informations soient détectées et référencées, l'information est déjà partagée un grand nombre de fois sur les réseaux sociaux et il devient très difficile d'arrêter sa propagation. De plus, la majorité des utilisateurs des réseaux sociaux ignorent l'existence ou n'ont pas le réflexe d'utiliser ces ressources pour vérifier les publications qu'ils partagent. Il est donc essentiel que les

8. https://fiskkit.com/

utilisateurs des réseaux sociaux en ligne soient, avant tout, conscients des conséquences désastreuses de la prolifération des fausses nouvelles sur les individus et la société.

Pour cela, il est important d'éduquer et former les utilisateurs à ces nouvelles plateformes et à ces nouvelles possibilités d'accès à l'information. Les médias et l'école sont les acteurs majeurs de cette action éducative. Cela peut être fait en apprenant par exemple les réflexes de vérification des faits [9].

De leur côté, et en dépit des actions prises par les média sociaux pour identifier les fausses nouvelles et sensibiliser leurs utilisateurs, ces plateformes devraient jouer un rôle plus important dans cette perspective. À titre d'exemple, Facebook identifie les fausses nouvelles avec l'aide des utilisateurs et des vérificateurs de faits (figure 2.4). Les utilisateurs signalent à la plateforme d'éventuelles rumeurs, qui sont envoyées par la plateforme à des organisations de vérification des faits comme FactCheck et Snopes pour vérification. Les rumeurs vérifiées sont publiquement signalées comme contestées par un tiers chaque fois qu'elle apparaissent sur le réseau social. Les utilisateurs reçoivent un autre avertissement s'ils persistent à partager ces fausses informations avérées.

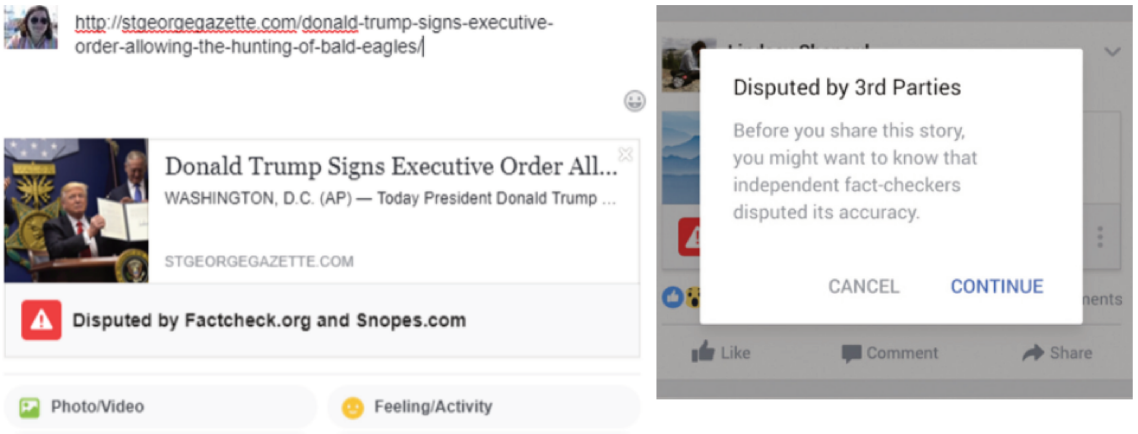
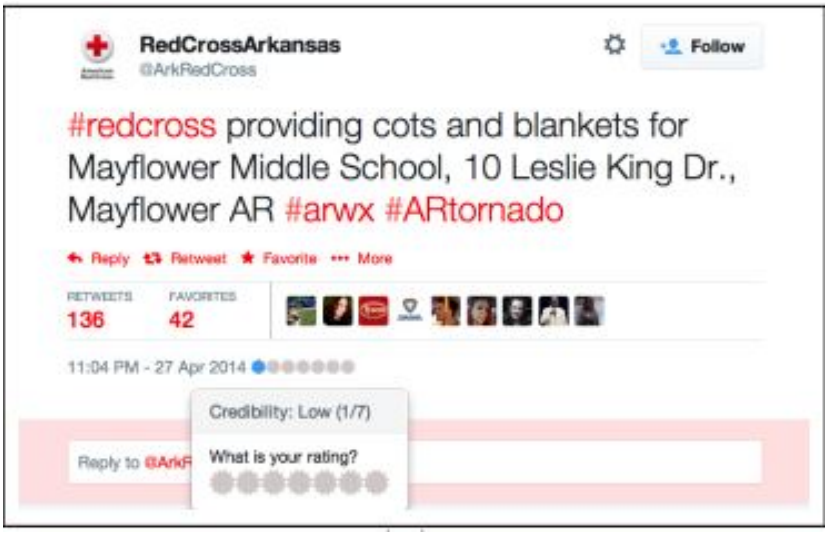


FIGURE 2.4 – Exemple de stratégie de lutte contre les rumeurs dans Facebook

Twitter utilise une stratégie semi-automatique combinant une évaluation automatique et l'annotation des utilisateurs pour signaler d'éventuels faux *tweets* (figure 2.5). Chaque *tweet* se voit attribuer une note de crédibilité générée automatiquement par un algorithme [Gupta et al., 2014]. Les utilisateurs peuvent donner leur avis s'ils ne sont pas d'accord avec l'évaluation (figure 2.5a). Dans ce cas, ils sont invités à indiquer ce qu'ils estiment être la note de crédibilité (figure 2.5b). Le retour d'information fourni par les utilisateurs est enregistré dans une base de données pour ré-entraîner le système.

9. https://www.lumni.fr/video/c-est-quoi-une-rumeur

(a)



(b)

FIGURE 2.5 – Exemple de stratégie de lutte contre les rumeurs dans Twitter

Enfin, pour tenter d'endiguer les informations mensongères ou trompeuses, le géant du Web Google a introduit en partenariat avec 115 structures de vérification de fait, dont 6 en France, une nouvelle rubrique *Fact-checking* sur son site d'informations [10] (figure 2.6). Désormais, dès lors qu'un internaute cherche une information qui a pu être vérifiée par l'une de ces structures, une mention apparaît en-dessous du résultat de la recherche. L'article du vérificateur est ainsi le premier résultat vu par l'internaute.

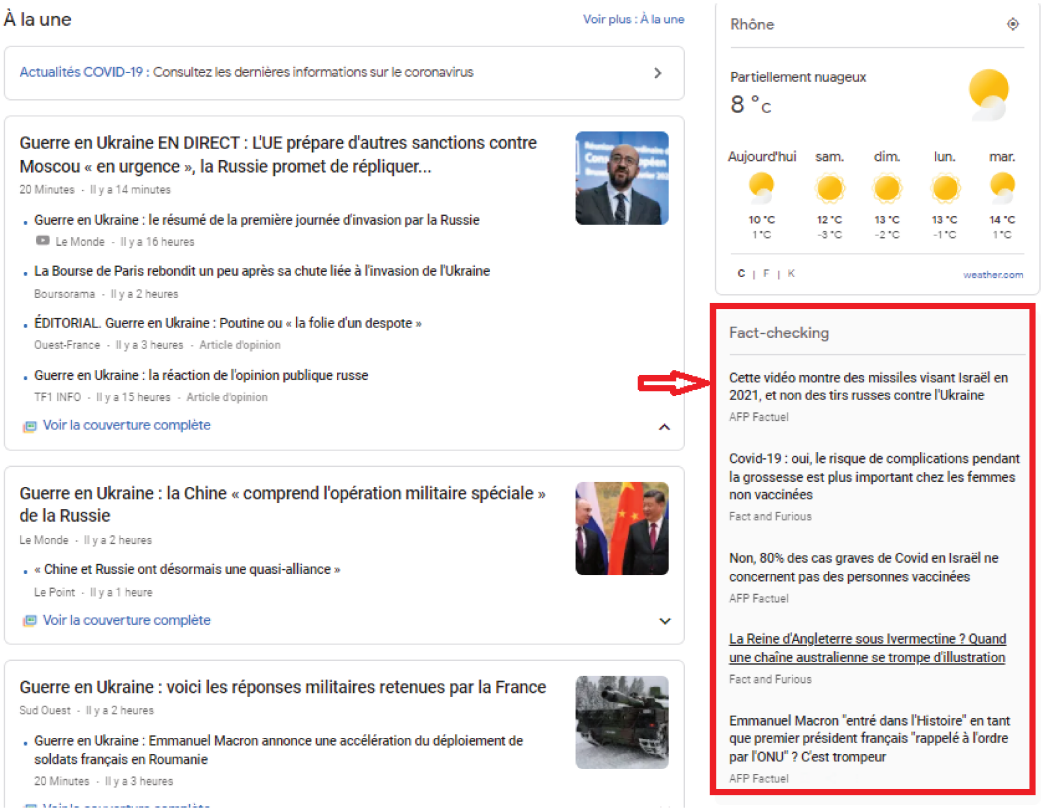


FIGURE 2.6 – Exemple de stratégie de lutte contre les rumeurs dans Google

Comme nous l'avons discuté dans le chapitre 1, les solutions manuelles proposées par le monde du journalisme demeurent difficiles, coûteuses et nécessitent des compétences spécifiques. Pour faire face à cela, dans un contexte notamment d'une quantité de données qui augmente de façon conséquente, plusieurs solutions automatiques ont été proposées par la communauté scientifique, elles sont détaillées dans la section suivante.

10. https://news.google.com/

2.3 Approches automatiques de détections de rumeurs

Les approches de détection des rumeurs sont diverses. Nous les catégorisons comme suit. Les méthodes guidées par les données (*data driven*) se concentrent sur les caractéristiques utilisées dans les travaux de la littérature pour présenter les rumeurs. Les approches guidées par le paradigme d'apprentissage utilisé se focalisent sur les familles d'algorithmes supervisés et non-supervisés. Nous traitons dans cet état de l'art les deux catégories ci-dessus. Comme illustré dans la figure 2.7, trois familles d'approches sont proposées, qui se basent sur l'extraction de caractéristiques efficaces pour représenter le contenu des rumeurs ou des algorithmes de prédiction robustes.

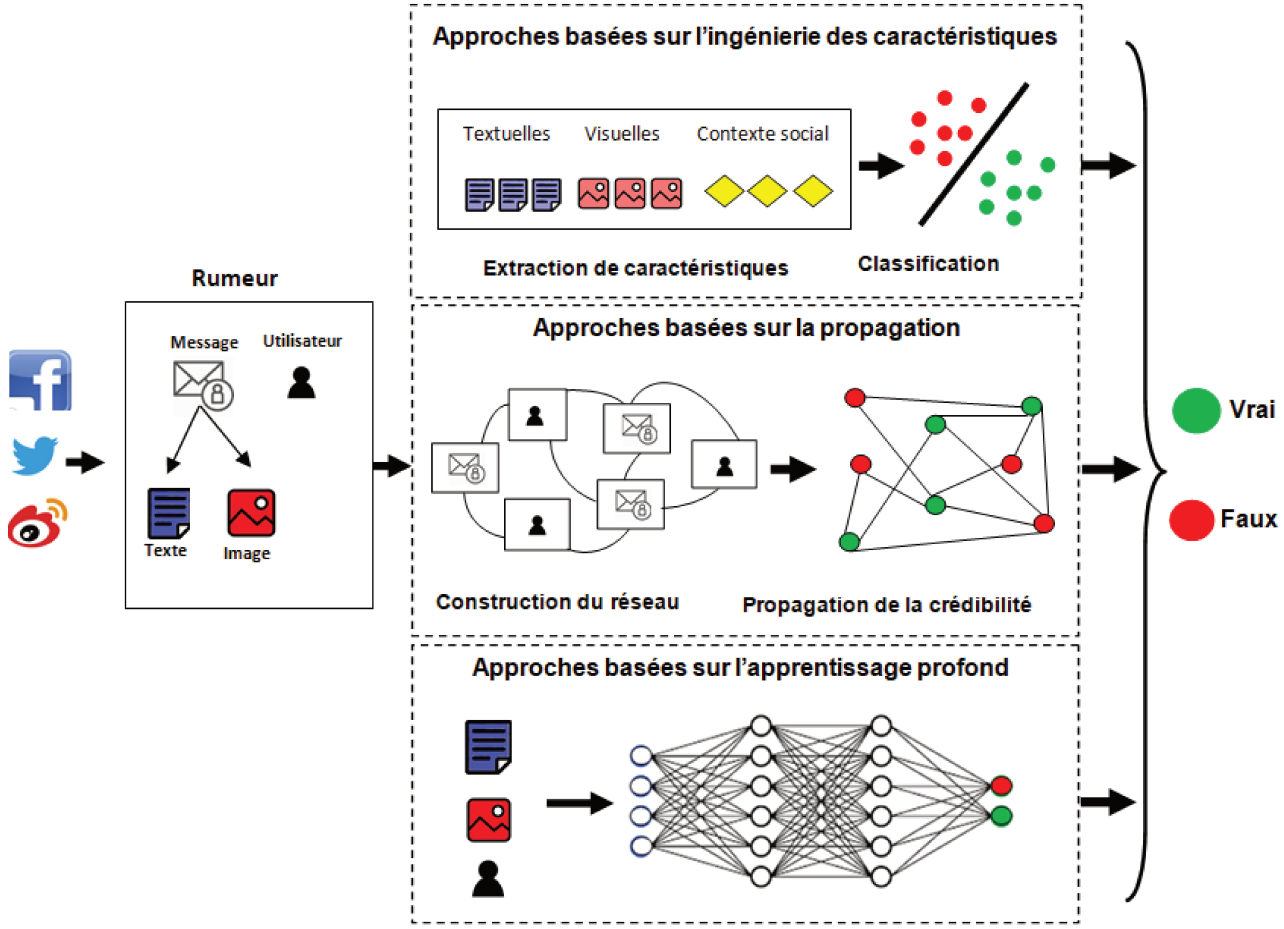


FIGURE 2.7 – Familles d'approches pour prédire la véracité des rumeurs

2.3.1 Approches basées sur l'ingénierie des caractéristiques

La majorité des travaux basée sur l'ingénierie des caractéristiques suivent une approche générale de fouille de données pour la détection des rumeurs qui comprend deux phases : (1) extraction de caractéristiques et (2) construction du modèle. La phase d'extraction de caractéristiques vise à représenter le contenu du message et les informations connexes dans une structure mathématique formelle, tandis que la phase de construction du modèle permet de construire des modèles d'apprentissage supervisé pour mieux différencier les fausses et vraies informations en se basant sur les représentations des caractéristiques. Les caractéristiques peuvent être extraites principalement à partir du contenu des messages et du contexte social.

2.3.1.1 Caractéristiques de contenu

Généralement, le contenu des nouvelles est principalement composé de données textuelles et visuelles. Deux catégories de caractéristiques textuelles générales sont couramment utilisées : (1) les caractéristiques linguistiques et syntaxiques et (2) les caractéristiques de style.

Caractéristiques linguistiques et syntaxiques : Ces caractéristiques se réfèrent à la composante fondamentale, à la structure et à la sémantique du langage naturel. Les caractéristiques linguistiques et syntaxiques sont toujours des sources précieuses pour l'analyse des rumeurs. Elles peuvent être extraites au niveau des caractères, des mots, des phrases et de l'événement associé.

Les caractéristiques extraites des caractères et des mots sont généralement calculées manuellement ou en utilisant les modèles les plus communément utilisés pour le traitement du langage naturel, comme les sacs de mots (*Bag of Words* – BoW), n-grammes, fréquence des termes (*Term Frequency* – TF) et fréquence des termes-fréquence inverse des documents (*Term Frequency-Inverse Document Frequency* – TF-IDF).

Un des premiers travaux utilisant ce type de caractéristiques classe les tweets comme « crédibles » ou « non-crédibles » en définissant plusieurs caractéristiques de comptage, comme le nombre total de mots et de caractères, le nombre de mots distincts et la longueur moyenne des mots dans une rumeur [Castillo et al., 2011]. Certaines

caractéristiques lexicales se sont révélées pertinentes pour la prédiction de la crédibilité, comme « le *tweet* contient une URL » et « le *tweet* contient un point d'interrogation ». Kwon et al. [2013] proposent d'autres caractéristiques linguistiques pour classer les messages comme vrai ou faux. Ils recherchent notamment la présence des pronoms de la première, deuxième et troisième personne (je, tu, il, elle). Les caractéristiques proposées permettent ainsi d'améliorer la prédiction par rapport au travail proposé par Castillo et al. [2011].

Zhao et al. [2015] partent de l'idée que les rumeurs provoquent des messages d'utilisateurs sceptiques, qui posent des questions ou se renseignent sur la véracité de la rumeur. Ils détectent les *tweets* de ce type en utilisant une sélection supervisée de caractéristiques sur un ensemble de messages labélisés. Les auteurs créent une liste manuelle de cinq expressions régulières (par exemple, « is (that — this — it) true ») qui sont utilisées pour identifier les *tweets* interrogatifs. Ils extraient ensuite les caractéristiques lexicales des *tweets* en utilisant les modèles unigrammes, bigrammes et trigrammes, ainsi que TF. Les tests du Chi-2 et du ratio de gain d'information sont utilisés pour sélectionner les caractéristiques pertinentes. Cette liste de caractéristiques est ensuite étudiée par des experts humains qui sélectionnent des phrases indépendantes des événements comme modèles lexicaux finaux pour les rumeurs. L'utilisation des experts humains renforce la cohérence des modèles lexicaux trouvés.

Les mots exprimant des sentiments ou des sémantiques spécifiques sont également des indices très importants pour caractériser le texte. Les marques émotionnelles (points d'interrogation et d'exclamation) et les émoticônes sont considérés comme des caractéristiques textuelles [Castillo et al., 2011]. L'émotion véhiculée dans le texte est aussi étudiée par Kwon et al. [2013], qui proposent de nombreuses caractéristiques linguistiques liées aux sentiments en se basant sur des dictionnaires de sentiment. Concrètement, ils utilisent un outil de sentiment appelé *Linguistic Inquiry and Word Count* (LIWC) pour compter les mots dans des catégories psychologiquement significatives. Après une étude comparative de ces caractéristiques, ils constatent que certaines catégories de sentiments sont distinctives pour la détection des rumeurs. Par exemple, les rumeurs sont significativement moins susceptibles de contenir des mots à effet positif (*love, nice, sweet*). Les utilisateurs sont aussi beaucoup plus susceptibles de mentionner des mots de négation (*no, not, never*) dans leur phrase et d'effectuer une action cognitive (*cause, know*) au contenu lié à la rumeur.

La présence de mots stylistiques peut également être utilisée pour la détection de rumeurs. Horne and Adali [2017] proposent un ensemble de caractéristiques comme

les mots vides (*stop words*), les signes de ponctuation, les guillemets, les négations (non, jamais, pas, etc.), les mots informels, les interrogations (comment, quand, quoi, pourquoi), les noms, les pronoms possessifs, les déterminants, les interjections, les verbes, les termes d'argot en ligne (tels que LOL ou BRB).

Les caractéristiques extraites au niveau de la phrase font référence à tous les attributs importants à l'échelle de la phrase. Ce sont des statistiques simples d'un message de rumeur, comme le nombre de mots-clés, le score de sentiment ou la polarité de la phrase. Pour représenter un document basé sur les mots qu'il contient, Ma et al. [2017] utilisent le modèle linguistique BoW, où chaque document est représenté comme un vecteur à v dimensions, où v est la taille du dictionnaire du corpus. Chaque élément du vecteur représente le score TF-IDF du mot correspondant dans le texte. Vicario et al. [2019] proposent d'autres caractéristiques, comme la polarité moyenne de la phrase (positive, neutre ou négative), la longueur moyenne de la phrase et la complexité de la phrase pour analyser les rumeurs. Les caractéristiques syntaxiques comme l'étiquetage morpho-syntaxique (*Part-of-Speech tagging* – PoS) sont également exploitées pour saisir les caractéristiques linguistiques des textes. Concrètement, chaque mot de la phrase est étiqueté sur la base de sa fonction syntaxique, telle que les noms, les pronoms et les adjectifs. Plusieurs travaux ont montré que la distribution de fréquence de ces étiquettes est étroitement liée au genre du texte considéré. Par exemple, les consultations médicales, les réunions de comité et les sermons ont chacun leur propre modèle distinctif. Horne and Adali [2017] et Ott et al. [2011] affirment que cette variation dans la distribution des balises PoS existe aussi par rapport à la véracité du texte des rumeurs.

Les modèles linguistiques simples utilisés pour représenter le texte brut souffrent de certaines lacunes. Par exemple, le modèle n-grammes est extrêmement épars et ne peut pas interpréter des échantillons de texte contenant des *tokens* inconnus. Le modèle BoW peut perdre des informations importantes en ignorant le contexte et la sémantique des mots. À cet effet, des travaux récents utilisent des modèles de plongement de mots (*word embedding*) comme Word2Vec, qui représente les mots en se basant sur leur similarité sémantique [Jin et al., 2017b].

Les caractéristiques extraites au niveau de l'événement du message visent à comprendre les relations sous-jacentes entre les messages au sein d'un corpus. Wu et al. [2015] définissent un ensemble de caractéristiques de sujets pour la détection des rumeurs sur le réseau social chinois Sina Weibo. Ils utilisent le modèle *Latent Drichlet Allocation* (LDA) avec une distribution de 18 sujets sur tous les messages. Chaque message peut appartenir à un ou plusieurs sujets. Ils transforment le vecteur de distribution à 18

dimensions en vecteur binaire en fixant les k sujets les plus probables à 1 et les autres à 0. La valeur de k est fixée par les auteurs. Jin et al. [2015] regroupent les sujets en fonction du thème auquel un message fait référence et extraient des caractéristiques à la fois au niveau du message et au niveau du sujet. Ils supposent que les messages relevant d'un même sujet ont probablement des valeurs de crédibilité similaires. Sous cette hypothèse, ils regroupent les messages en différents sujets et obtiennent les caractéristiques au niveau du sujet en agrégeant les caractéristiques au niveau du message (par exemple, moyenne des valeurs d'une même caractéristique pour tous les messages). Ils affirment que ce type de caractéristiques au niveau du sujet peut réduire l'impact des données bruitées tout en conservant la plupart des détails au niveau du message.

Caractéristiques de style : Elle sont spécifiquement conçues pour capturer les indices trompeurs dans les styles d'écriture afin de distinguer les rumeurs. Les utilisateurs malveillants expriment des informations trompeuses en cachant intentionnellement leur style d'écriture ou en tentant d'imiter d'autres utilisateurs. En essayant de capturer les caractéristiques distinctives des styles d'écriture entre les utilisateurs légitimes et les comptes anormaux, l'analyse basée sur le style joue un rôle important dans l'identification des rumeurs en ligne. Ahmed [2017] propose deux types de caractéristiques. (1) Les caractéristiques d'édition représentent les actions de l'utilisateur lors de l'édition du texte. Il s'agit notamment des touches « Retour arrière » et « Supprimer », des touches fléchées du clavier et du nombre de fois où la souris est utilisée. (2) Les caractéristiques de durée peuvent par exemple être la durée moyenne du mot, l'intervalle moyen entre les mots, etc. Ahmed [2017] constate que les faux créateurs de contenu ont besoin de plus de temps pour finir d'écrire et qu'ils ont tendance à faire plus d'erreurs. Castillo et al. [2011], Horne and Adali [2017] et Jin et al. [2016b] exploitent de nombreuses caractéristiques pour l'analyse du style d'écriture, comme la fraction de messages qui contiennent des liens externes, des mentions d'utilisateurs, des *hashtags*, des mots en majuscules pendant une période de temps. Afroz et al. [2012] proposent d'utiliser le domaine de la stylométrie pour décrire les propriétés stylistiques du texte. Ils considèrent que certaines caractéristiques linguistiques changent lorsque les auteurs cachent leur style d'écriture et, donc, l'identification de ces caractéristiques permettent de reconnaître la déformation stylistique. Ils prorosent une méthode permettant de distinguer la déformation stylistique de l'écriture normale, en utilisant un grand nombre de caractéristiques textuelles et contextuelles.

Caractéristiques visuelles : Contrairement au grand nombre de travaux qui exploitent le contenu textuel pour la classification des rumeurs, le contenu visuel, notamment les images, est peu exploré. Seuls quelques travaux très récents tentent d'extraire des caractéristiques du contenu visuel des messages contenant des rumeurs.

Gupta et al. [2013] font un premier effort pour comprendre la diffusion de fausses images sur Twitter pendant l'ouragan Sandy. Ils proposent un modèle de classification pour identifier les fausses images. Ils affirment que 86 % des *tweets* diffusant des fausses images sont des *retweets*. Cependant, leur travail ignore le contenu des image. Il se base principalement sur des caractéristiques textuelles extraites du contenu des *tweets* et des profils des utilisateurs associés. Gupta et al. [2012] définissent une caractéristique permettant de noter si l'utilisateur a une image de profil afin d'évaluer la crédibilité des utilisateurs. Wu et al. [2015] utilisent une caractéristique *has multimedia* pour indiquer la présence d'un contenu multimédia attaché à un *tweet* (si le *tweet* est accompagné d'une image, d'une vidéo ou d'un fichier audio). Lors d'une investigation relative à la perception des utilisateurs concernant la crédibilité du contenu sur Twitter, Morris et al. [2012] ont découvert que les indicateurs importants sur lesquels les utilisateurs jugent la crédibilité sont les informations visibles au premier regard, notamment celles relatives au profil de l'utilisateur (son nom et son image), qui ont un grand impact sur la crédibilité des messages publiés par cet utilisateur. Pour prédire automatiquement si un *tweet* qui partage un contenu multimédia est faux ou vrai, Boididou et al. [2015] proposent la tâche *Verifying Multimedia Use* (VMU). Des caractéristiques textuelles et de forensique extraites du contenu visuel des images attachées aux *tweets* (par exemple, carte de probabilité de la double compression JPEG alignée, *Photo-Response Non-Uniformity*, etc.) sont utilisées comme éléments de base pour cette tâche. Cependant, les auteurs concluent dans une autre étude que l'utilisation de caractéristiques de forensique ne conduit pas à une amélioration notable de la performance, car le traitement automatique du contenu multimédia attaché aux publications Twitter supprime l'essentiel des traces pertinentes du forensique des images [Boididou et al., 2018].

2.3.1.2 Caractéristiques du contexte social

Ces caractéristiques sont conçues pour refléter le motif de distribution des informations et l'interaction entre les utilisateurs en ligne. On peut lister trois types d'interactions sociales sur les médias sociaux :

— les interactions entre les utilisateurs, telles qu'ajouter un ami (*friend*) et suivre (*follow*) ;
— les liens entre les contenus multimédias sont formés par le biais de balises, de *hashtags* ou de liens URL ;
— les interactions entre les utilisateurs et le contenu, comme poster un message, ajouter un commentaire ou partager une publication.

Les trois principaux types de caractéristiques sociales sont les caractéristiques basées sur l'utilisateur, les caractéristiques de propagation et les caractéristiques temporelles.

Les caractéristiques utilisateur visent à décrire le profil d'un utilisateur unique ou d'un groupe d'utilisateurs. Les caractéristiques au niveau individuel sont extraites pour évaluer la crédibilité et la fiabilité de chaque utilisateur en utilisant divers aspects, tels que l'âge du compte sur le *microblog*, le nombre de *followers/followees*, le nombre de *tweets* postés [Castillo et al., 2011]. Les caractéristiques de groupe capturent les profils généraux des groupes d'utilisateurs. L'hypothèse est que les diffuseurs de fausses et de vraies informations peuvent former des communautés différentes avec des caractéristiques similaires, qui peuvent être décrites au niveau du groupe [Yang et al., 2012]. Les caractéristiques de niveau groupe couramment utilisées proviennent de l'agrégation (par exemple, en faisant la moyenne et en pondérant) des caractéristiques de niveau individuel, comme le pourcentage d'utilisateurs vérifiés et le nombre moyen de *followers* [Kwon et al., 2013].

Les caractéristiques de propagation sont extraites en construisant des réseaux spécifiques entre les utilisateurs qui publient des messages connexes sur les médias sociaux. Castillo et al. [2011] proposent quelques statistiques à partir des arbres de propagation qui peuvent être construits à partir des *retweets* des messages, telles que la profondeur ou la taille moyenne de l'arbre de propagation. Kwon et al. [2013] les étendent à 15 caractéristiques structurelles extraites à partir, premièrement, du réseau de diffusion qui retrace la trajectoire de la diffusion de la rumeur, où les nœuds représentent les utilisateurs et les arêtes représentent les chemins de diffusion de l'information entre eux. Plus précisément, un chemin de diffusion entre deux utilisateurs u_i et u_j existe si et seulement si u_i suit u_j et u_j publie un message sur une rumeur uniquement après que u_i le fait. Deuxièmement, un réseau d'amitié indique la structure *following/followee* des utilisateurs qui publient des *tweets* connexes. Ces caractéristiques concernent notamment le nombre de nœuds et de liens, le degré médian et la densité de ces réseaux.

Enfin, les caractéristiques temporelles marquent les points de temps importants ou

le cycle de vie de la diffusion des rumeurs. Ma et al. [2015] proposent une méthode permettant de discrétiser le flux temporel et de capturer la variation des caractéristiques temporelles. Giasemidis et al. [2016] découpent chaque événement de rumeur en 20 intervalles de temps et extraient des caractéristiques pour chaque sous-ensemble de messages. Kwon et al. [2017] étudient la stabilité des caractéristiques dans le temps et constatent que, pour la détection des rumeurs, les caractéristiques linguistiques et celles liées à l'utilisateur sont de bons indicateurs durant la phase initiale de la propagation, tandis que les caractéristiques structurelles et temporelles ont tendance à avoir de bonnes performances à long terme.

Jusqu'ici, nous avons introduit dans cette section les différentes caractéristiques extraites du contenu des publications et du contexte social. La deuxième phase dans cette approche basée sur l'ingénierie des caractéristiques est d'utiliser des modèles d'apprentissage automatique pour la détection des rumeurs. Avec un nombre suffisant de caractéristiques disponibles, de nombreuses méthodes de classification supervisée sont proposées dans la littérature. Ces méthodes doivent être capables d'intégrer et d'interpréter ces caractéristiques pour une détection robuste des rumeurs. La plupart des travaux ont expérimenté plus d'un algorithme de classification pour trouver le modèle le plus approprié, notamment les arbres de décision [Castillo et al., 2011; Ma et al., 2015; Giasemidis et al., 2016], les réseaux bayésiens [Castillo et al., 2011], les forêts aléatoires [Kwon et al., 2013; Jin et al., 2016b], la régression logistique [Giasemidis et al., 2016; Jin et al., 2016b] et les machines à vecteurs de support (*Support Vector Machines* – SVM) [Castillo et al., 2011; Kwon et al., 2013; Yang et al., 2012; Jin et al., 2016b; Ma et al., 2015].

Pour mieux agréger les différentes caractéristiques, certains travaux proposent de nouveaux algorithmes. Wu et al. [2015] proposent un SVM avec une technique hybride de noyau consistant en un noyau à marche aléatoire (*random walk kernel)* et un noyau de fonction de base radiale (*Radial Basis Function kernel* – RBF). Le noyau à marche aléatoire est spécifiquement conçu pour capturer les caractéristiques de propagation de l'arbre des commentaires d'un message, tandis que le noyau RBF est appliqué aux caractéristiques du contenu et de l'utilisateur. Considérant les rumeurs comme des anomalies, Chen et al. [2016] traitent la détection de rumeurs comme un problème de détection d'anomalies. Ils effectuent une analyse factorielle des données mixtes (*Factor Analysis of Mixed Data* – FAMD) sur les caractéristiques proposées pour détecter ces anomalies. Deux stratégies basées sur la distance euclidienne et la similarité cosinus sont proposées pour décrire le degré de déviation.

2.3.2 Approches basées sur la propagation

Ces approches exploitent le cheminement du message, qui correspond à son historique, soit toutes les personnes ayant partagé ce message depuis son auteur originel jusqu'à l'utilisateur associé au message à traiter.

Les approches basées sur l'ingénierie des caractéristiques évaluent chaque message et événement individuellement. Cependant, il existe certaines corrélations sous-jacentes entre les messages et les événements dans les médias sociaux. Une observation simple est que les messages similaires ont tendance à avoir la même véracité dans un événement. Les approches basées sur la propagation explorent des relations entre les entités (événements, messages, utilisateurs) et évaluent la véracité des messages et des événements dans leur ensemble. Les approches de classification des rumeurs basées sur la propagation de la crédibilité comportent généralement deux étapes principales.

1. Construction du réseau de crédibilité : les entités impliquées dans la détection des rumeurs, telles que les messages, les utilisateurs ou les événements, sont définies comme des nœuds du réseau. Chaque nœud a une valeur de crédibilité initiale. Les liens entre ces entités sont définis et calculés en fonction de leur relation sémantique ou de leur relation d'interaction dans le média social.
2. Propagation de la crédibilité : sous certaines hypothèses de cohérence des nœuds et de régularité du réseau, les valeurs de crédibilité sont propagées sur le réseau construit selon des liens pondérés jusqu'à la convergence ce qui donne l'évaluation finale de la crédibilité pour chaque entité.

Le problème de la propagation est formé comme une tâche d'apprentissage de graphe semi-supervisée, où certains sommets sont connus comme crédibles ou non, et dont le but est d'estimer la crédibilité des autres sommets [Zhu and Ghahramani, 2002; Zhu et al., 2003]. Par rapport à la classification directe sur une entité individuelle, les approches basées sur la propagation peuvent tirer parti des relations entre entités et obtenir des résultats robustes.

Gupta et al. [2012] construisent un réseau composé d'utilisateurs, de messages et d'événements sous deux hypothèses : (1) les utilisateurs crédibles n'offrent en général pas de crédibilité aux événements de rumeurs ; (2) les liens entre les messages crédibles ont des poids plus importants que ceux des messages de rumeurs, car les messages dans un événement de rumeur ne font pas de déclarations cohérentes. Les valeurs de crédibilité

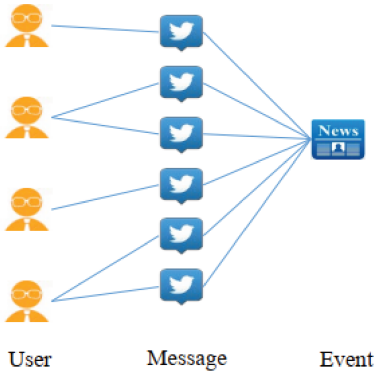
initiales de chaque message sont obtenues à partir des résultats d'un classifieur basé sur des caractéristiques similaires à celles introduites par Castillo et al. [2011] (section 2.3). Ils sont ensuite propagés sur ce réseau à l'aide d'itérations d'un algorithme de type PageRank [11].

Inspirés par l'idée de relier toutes les entités et l'exploitation des implications inter-entités pour la propagation de la crédibilité, Jin et al. [2014] proposent un réseau de crédibilité à trois couches construit à partir de différents niveaux sémantiques d'un événement : couche de message, couche de sous-événement et couche d'événement. Elles sont toutes basées sur le contenu et ont des relations directes avec la crédibilité des informations. Les sous-événements sont divers points de vue d'un même événement, qui sont des groupes de messages représentant des parties ou des sujets principaux d'un événement. Pour être plus précis, le réseau est construit comme suit : un message est lié à un sous-événement, ce dernier est lié à un événement, tous les messages sont liés entre eux, de même que les sous-événements. Sous l'hypothèse que les entités avec un grand poids de lien aient des valeurs de crédibilité similaires, le problème de propagation de crédibilité est formulé comme un problème d'optimisation de graphe.

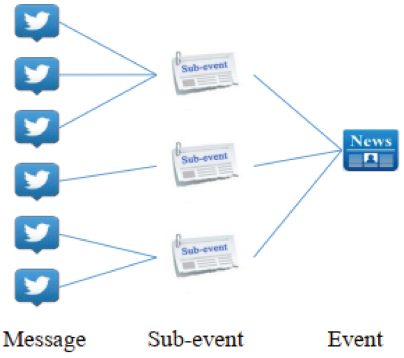
Selon Jin et al. [2016a], il existe deux types de relations entre les messages sur les *microblogs*. L'une est la relation de soutien, où les messages exprimant le même point de vue soutiennent la crédibilité des uns et des autres. L'autre relation est de s'opposer. Les messages expriment des points de vue contradictoires et réduisent la crédibilité de chacun d'entre eux. Partant de cette observation, les auteurs proposent un réseau de crédibilité avec des relations de soutien et d'opposition, en exploitant les points de vue contradictoires comme suit. (1) Les points de vue contradictoires sont extraits par une méthode de modèle thématique [Trabelsi and Zaiane, 2014]. Ce modèle représente chaque message comme un mélange d'événements et de points de vue pour chaque événement. Ces paires événement-point de vue sont ensuite regroupées sous un certain nombre de contraintes pour former les points de vue conflictuels finaux. (2) Tous les messages d'un événement sont liés les uns aux autres. Le poids du lien est calculé comme la distance entre les représentations de probabilité des deux messages du modèle thématique. La polarité du lien est définie à partir du regroupement des points de vue : les messages ayant les mêmes points de vue forment le lien positif, sinon ils forment le lien négatif. Similaire à celle de Jin et al. [2014], la propagation de la crédibilité dans ce réseau est également définie comme un problème d'optimisation de graphe. La figure 2.8 illustre la structure de ces trois réseaux de propagation.

11. https://fr.wikipedia.org/wiki/PageRank

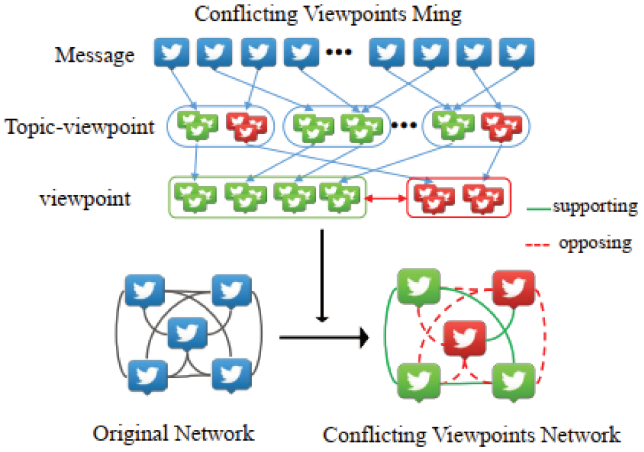
Les approches basées sur la propagation tirent parti des interrelations des messages crédibles des médias sociaux pour prédire la crédibilité des événements associés. Des réseaux de crédibilité homogènes et hétérogènes peuvent être construits pour le processus de propagation. Cependant, une limite présentée par ces approches est qu'elles ignorent l'information textuelle représentée par les rumeurs.



(a) Réseau proposé par Gupta et al. [2012]



(b) Réseau proposé par Jin et al. [2014]



(c) Réseau proposé par Jin et al. [2016a]

FIGURE 2.8 – Réseaux pour la détection de rumeurs basée sur la propagation de la crédibilité

2.3.3 Approches basées sur l'apprentissage profond

Contrairement aux modèles basés sur l'apprentissage classique, qui dépendent de caractéristiques élaborées manuellement, les modèles profonds sont capables d'extraire

automatiquement des caractéristiques et des représentations cachées dans le texte et les images. Les travaux de ces approches utilisent essentiellement deux structures de réseaux de neurones, les réseaux de neurones récurrents (*Recurrent Neural Networks* – RNN) qui modélisent les données de la rumeur comme des données séquentielles [Ma et al., 2015; Chen et al., 2018], et les réseaux de neurones convolutifs (*Convolutional Neural Network* – CNN) pour capturer des caractéristiques locales et globales.

Ma et al. [2016] observent qu'un événement de rumeur se compose d'un message original et d'un groupe de messages connexes, y compris les réponses et les commentaires, qui créent un flux continu de messages. Ainsi, ils modélisent les données de rumeur comme une série temporelle de longueur variable. Cependant, un événement de rumeur se compose de dizaines de milliers de messages. Les auteurs regroupent donc les messages dans des intervalles de temps et les traitent comme une seule unité dans une série temporelle qui est ensuite modélisée en utilisant une séquence RNN. Dans chaque intervalle, ils utilisent les valeurs TF-IDF des k premiers termes du vocabulaire comme entrée. Leur modèle vise à apprendre des représentations temporelles et textuelles à partir des données de rumeurs. Leurs expériences démontrent que leur modèle surclassent les travaux utilisant l'ingénierie des caractéristiques textuelles.

Certains mots malveillants dans le contenu peuvent être fortement liés à la catégorie de rumeur. Pour mieux comprendre les mots auxquels le modèle prête le plus d'attention, Chen et al. [2018] utilisent un mécanisme d'attention basé sur un réseau RNN. L'une des hypothèses de leur travail est que les caractéristiques textuelles des rumeurs peuvent changer d'importance avec le temps et qu'il est crucial de déterminer lesquelles sont les plus importantes pour la tâche de détection. Les auteurs regroupent d'abord les messages par intervalles de temps. À chaque pas temporel, l'état caché du RNN se voit attribuer un paramètre de pondération pour mesurer son importance et sa contribution aux résultats. Les performances expérimentales démontrent l'efficacité du mécanisme d'attention et que la majorité des mots liés à l'événement lui-même sont moins utilisés que les mots exprimant le doute et la colère des utilisateurs causés par les rumeurs.

Ma et al. [2018] proposent un modèle de détection des rumeurs basé sur des réseaux neuronaux récursifs. L'entrée de leur modèle est un arbre de propagation dont la racine est un message source, et chaque nœud de l'arbre peut être un message de réponse. La sémantique du contenu des messages et les relations entre eux peuvent être capturés conjointement via le processus d'apprentissage récursif des caractéristiques tout au long de la structure arborescente. Ce modèle atteint des performances bien meilleures que les approches de l'état de l'art de l'époque et démontre une capacité supérieure à détecter

les rumeurs au début de la propagation.

Ruchansky et al. [2017] se concentrent sur trois caractéristiques des rumeurs : le texte d'un article, la réponse des utilisateurs qu'il reçoit et la source des utilisateurs qui en font la diffusion. Ainsi, ils proposent un modèle hybride qui combine les trois caractéristiques pour une prédiction automatique et plus précise. Le modèle est composé de trois modules : *Capture, Score* et *Integrate* (figure 2.9). Le premier module est basé sur la réponse et le texte. Il utilise un réseau RNN pour capturer le profil temporel de l'activité des utilisateurs sur un article donné. Le deuxième module apprend la caractéristique de la source en se basant sur le comportement des utilisateurs. Un utilisateur est représenté par un vecteur. Dans le troisième module, les résultats des deux premiers modules sont intégrés dans un vecteur qui est utilisé pour la classification.

Nguyen et al. [2017] se concentrent sur la détection au début de la propagation de la rumeur. Ils proposent l'utilisation d'un CNN et d'un RNN. Le CNN est utilisé pour extraire une séquence de représentations de phrases de haut niveau afin d'apprendre les représentions cachées de chaque *tweet* liées à des rumeurs et ainsi prédire la véracité de chaque *tweet*. Ensuite, le RNN est utilisé pour analyser les séries temporelles (séries de prédiction au niveau des publications) produites par le CNN pour obtenir la prédiction finale.

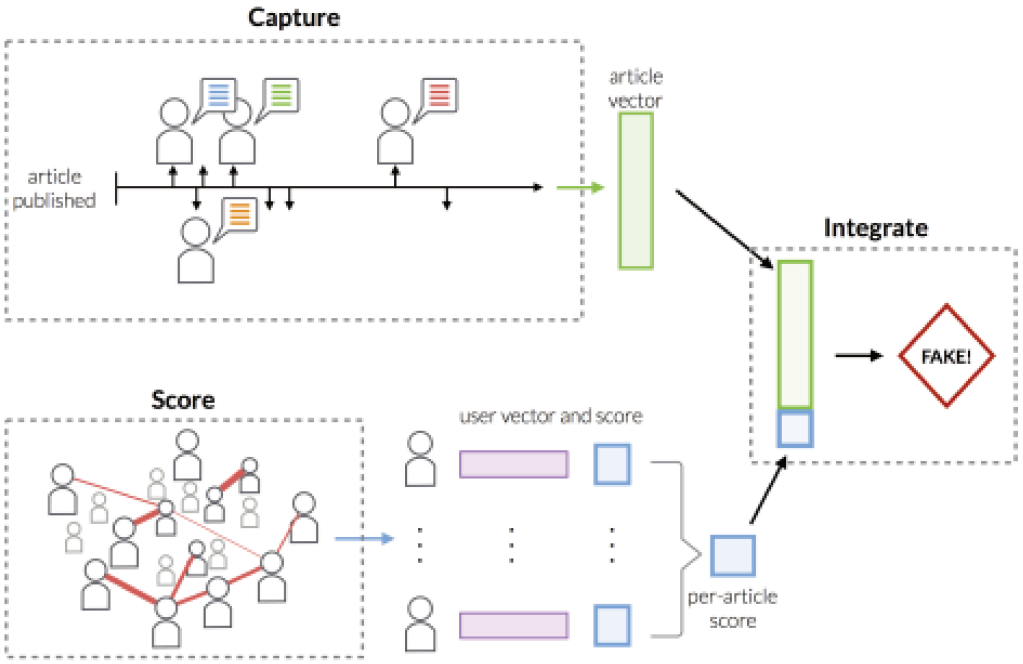


FIGURE 2.9 – Illustration de l'approche proposée par [Ruchansky et al., 2017]

2.3.4 Discussion

La présentation des différentes approches de détection de rumeurs permet de tirer la synthèse et les conclusions suivantes.

Les méthodes basées sur l'ingénierie des caractéristiques utilisent ces caractéristiques pour décrire la distribution des rumeurs dans un espace à haute dimension. L'hyperplan de séparation entre les classes est appris par des classifieurs d'apprentissage automatique traditionnel. La phase d'ingénierie des caractéristiques est une étape cruciale dans le processus d'analyse des rumeurs. Elle permet d'intégrer divers types d'informations dans le processus d'apprentissage. Un autre avantage de ces approches est qu'elles sont en mesure de procurer certains éléments d'explicabilité et d'interprétabilité quant aux décisions prises. Nous pensons que de telles explications sont nécessaires, particulièrement dans le contexte des rumeurs où la vie privée des personnes est en jeu. Cependant, dans le cas où la rumeur manque de certaines caractéristiques discriminantes, ces méthodes conduisent souvent à des résultats instables.

Les méthodes basées sur la propagation de la crédibilité utilisent la structure hétérogène du réseau social. Les messages et les utilisateurs sont reliés dans un réseau entier par des méthodes d'optimisation basées sur les graphes et leurs crédibilités sont évaluées dans leur ensemble. Cependant, il est évident que ces travaux ignorent une composante importante de la rumeur, à savoir l'information textuelle contenue dans les messages.

Les approches basées sur l'apprentissage profond exploitent la capacité des différentes structures de réseaux de neurones pour extraire automatiquement et apprendre des caractéristiques complexes des rumeurs. Comparées aux deux familles d'approches précédentes, celles-ci améliorent considérablement les performances de la prédiction. Cependant, l'inconvénient majeur de ces techniques est le manque d'éléments d'explicabilité et d'interprétabilité dans les résultats de classification.

Il est à noter que les travaux s'inscrivant dans cette démarche de prédiction de la véracité des rumeurs ne se basent pas sur la véracité des images elles-mêmes. C'est précisément ce point que nous détaillons dans la section suivante.

2.4 Détection de modifications dans une image numérique

L'analyse de la véracité des images est une tâche difficile qui nécessite de relever de multiples défis du fait du large panel de possibilités quant aux types de modifications possibles. La catégorisation des techniques de détection de modifications dans une image est présentée dans cette section en suivant la classification la plus courante de l'état de l'art. Il est à noter que ces techniques ne s'inscrivent pas spécialement dans une démarche liée à la vérification de rumeurs qui constitue notre objectif. Pour cette raison, nous ne détaillons pas la présentation de ces techniques, le domaine étant extrêmement large. Notre objectif est plutôt de donner un aperçu pertinent pour nos travaux.

Les images numériques peuvent être altérées en utilisant soit des méthodes passives, soit des méthodes actives. La figure 2.10 présente la cartographie de ces méthodes, présentées successivement dans la suite de cette section.

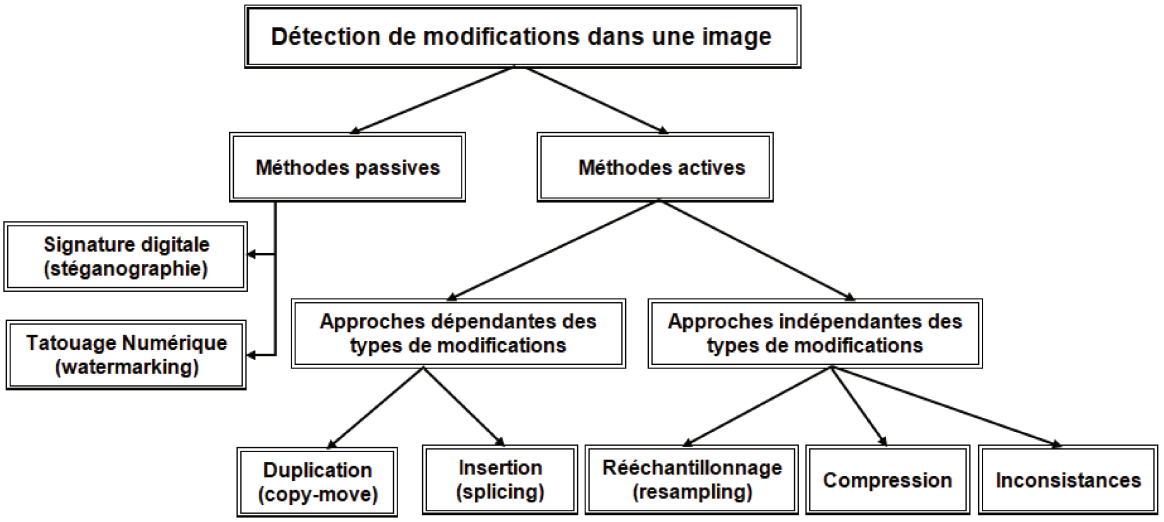


FIGURE 2.10 – Cartographie des méthodes de détection de modifications d'image

2.4.1 Méthodes actives

Les méthodes actives exploitent certaines informations insérées dans l'image par le dispositif de prise de vue pendant l'acquisition de l'image. Les données insérées dans l'image sont utilisées pour détecter l'origine de l'image ou pour constater une altération dans ladite image. Le tatouage numérique (*watermaking*) et les signatures

numériques (stéganographie) sont les deux principales catégories de techniques actives. Les modifications apportées par ces deux techniques ont la particularité d'être invisibles à l'œil nu.

Le tatouage numérique permet d'insérer des informations de copyright ou d'identification dans une image au moment de la capture. Pour vérifier la véracité de l'image, cet identifiant est comparé avec l'identifiant original. En cas de non correspondance, cela indique que l'image a été modifiée après le processus d'acquisition. Le tatouage est un mécanisme efficace pour préserver l'intégrité de l'image numérique, mais plusieurs difficultés rendent son utilisation peu pratique. À titre d'exemple, il n'existe que quelques appareils ou caméras qui ont la capacité d'intégrer un filigrane lors de l'acquisition d'images.

La stéganographie permet de cacher un message dans une image numérique. Ce type de techniques modifie les valeurs de quelques pixels afin de cacher un message dans l'image sans modifier l'aspect visuel de l'image.

2.4.2 Méthodes passives

Ces méthodes analysent le contenu et la structure d'une image pour en évaluer l'authenticité ou l'intégrité, sans supposer la présence d'une signature intégrée dans l'image originale. Les modifications numériques ne laissent pratiquement aucun indice visuel d'altération, mais peuvent perturber la structure de l'image, entraînant la modification des caractéristiques de l'image (l'éclairage, les ombres, etc.), ce qui les rend incohérentes. Ces incohérences peuvent être utilisées pour détecter la modification. On distingue deux types de méthodes passives en fonction de la dépendance du type de modification ou pas.

2.4.2.1 Approches indépendantes du type de modifications

Ces techniques recherchent des changements intrinsèques dans les caractéristiques de l'image et ne visent pas un type de modification spécifique. Elles se décomposent en trois catégories : les techniques basées sur la compression, la détection d'inconsistances et la détection de ré-échantillonnage. Ce dernier type correspond à plusieurs définitions dans l'état de l'art. Nous retenons celle qui correspond à l'ensemble des modifications

apportées à l'image entière, telles que la rotation, la translation, le changement de couleurs, etc.

Les techniques basées sur la compression utilisent les particularités de différents formats d'images pour détecter des incohérences dans les valeurs des pixels. Le format *Joint Photographic Experts Group* (JPEG) est le plus étudié, car c'est le standard de compression d'images le plus populaire. Ce format est aussi capable de compresser l'image sans mémoriser l'intégralité des données (perte de données engendrée par la compression). Plusieurs travaux développent des techniques pour savoir si une image est doublement compressée ou non [Bianchi and Piva, 2012; Li et al., 2018; Huang et al., 2018]. Le principe est qu'une image est compressée la première fois lors de la prise de la photo, puis une seconde fois au moment de l'enregistrement si l'image vient d'être modifiée. Ainsi, la partie non altérée de l'image est donc compressée deux fois, alors que la région altérée ne semble l'être qu'une seule fois, car l'artefact JPEG introduit par la compression initiale a été détruit par la modification. Ces techniques nécessitent de travailler avec un format particulier d'images, ce qui n'est pas le cas général des images circulant sur les réseaux sociaux, car le format d'image peut être changé par le réseau social lui même lors de la soumission de la publication.

Une seconde catégorie de techniques indépendantes du type d'attaque est basée sur les inconstances dans l'image. Ces approches se focalisent sur la détection d'aberrations chromatiques et d'incohérences de la lumière. Johnson and Farid [2006] implémentent un modèle pour l'aberration chromatique latérale et une technique automatique pour estimer les paramètres de ce modèle, qui est basé sur la maximisation de l'information mutuelle entre les canaux de couleur. Cependant, cette approche est efficace lorsque la région manipulée est relativement petite. Il est aussi observé que la détection est peu performante sur les images de faible qualité.

D'autres approches utilisent le signal *Photo Response Non-Uniformity* (PRNU) qui est un signal apposé par un appareil de capture sur une image lors de la prise de la photo. En cas de modification de l'image, ce signal global à l'image est perturbé dans la zone modifiée [Goljan et al., 2010]. Cependant, ces approches se basent sur une connaissance *a priori* du type d'appareil photo utilisé pour la capture de la photo, et donc du signal apposé qui est propre à chaque modèle. Cozzolino et al. [2017] proposent une technique utilisant le PRNU tout en relaxant ces hypothèses, afin de se rapprocher d'un scénario du monde réel. Ils commencent par réaliser un regroupement (*clustering*) d'images provenant d'une base dont nous savons que les images d'un même regroupement proviennent d'un même appareil photo, et donc ayant le même signal

PRNU. Ils approximent le signal PRNU de l'image à analyser et cherchent le signal PRNU le plus similaire pour une image donnée. Le signal PRNU estimé de ce cluster est utilisé sur l'image requête. Toutefois, les auteurs n'indiquent pas la complexité de leur méthode.

2.4.2.2 Approches dépendantes du type de modifications

Les deux principaux types de ces approches sont l'épissage (*splicing*) et le copier-déplacer (*copy-move*). Dans une attaque par copier-déplacer, une partie d'une image est copiée et collée dans une autre partie de la même image. Un exemple de cette attaque est donné dans la figure 2.11, où une partie de l'image originale 2.11a est ajoutée afin d'obtenir l'image modifiée 2.11b.

(a) (b)

FIGURE 2.11 – Exemple d'image modifiée par copier-déplacer (à droite) avec l'image originale (à gauche)

Partant du principe que les pixels de la zone dupliquée sont en double dans l'image, l'objectif principal des méthodes de détection de ce type de modification est de chercher les régions similaires ou identiques de l'image. Deux techniques sont utilisées pour détecter les régions similaires, les techniques basées sur les blocs et celles basées sur les points clés. Dans les méthodes basées sur les blocs, des algorithmes d'extraction de caractéristiques sont appliqués à l'image segmentée. Les vecteurs de caractéristiques obtenus sont ensuite comparés à l'aide d'algorithmes de correspondance par blocs, tels que le tri lexicographique ou la distance euclidienne.

Dans les méthodes basées sur les points clés, la segmentation en blocs est supprimée de l'étape de prétraitement de l'image. Des caractéristiques locales distinctives, telles que les coins et les bords de l'image, sont extraites à l'aide d'algorithmes d'extraction de caractéristiques de points clés, comme *Scale-Invariant Feature Transform* (SIFT), *Speeded Up Robust Features* (SURF) et *Harris corner points* (détecteur de coins de Harris). Chaque caractéristique est définie par un ensemble de descripteurs extraits d'une région autour de la caractéristique. Ces descripteurs et les caractéristiques sont ensuite comparés grâce à des algorithmes de correspondance, comme KNN ($k = 2$) et le *clustering* pour trouver la région copiée. Ces méthodes sont sensibles aux régions lisses ou homogènes, où peu de points clés sont extraits de la région modifiée.

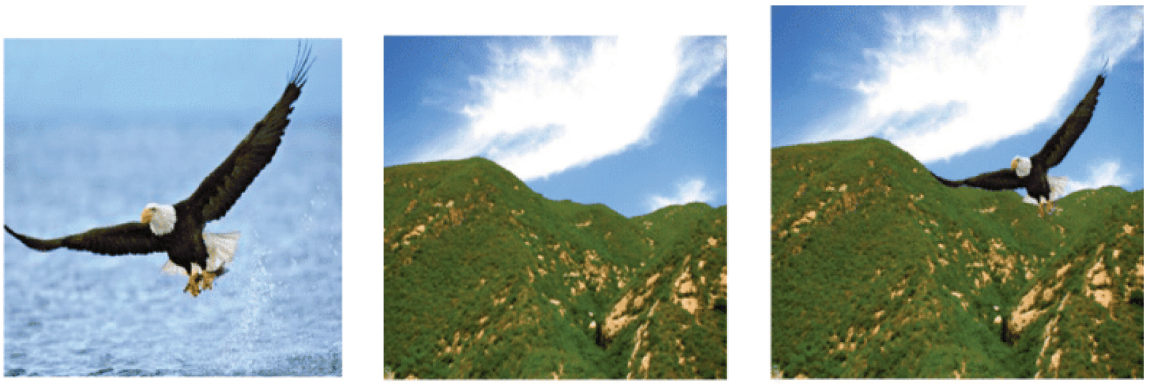
En général, les méthodes basées sur les blocs ont une complexité de calcul élevée en raison de l'appariement de grands nombres de blocs d'image. Cependant, les approches basées sur les points clés peuvent surmonter cet inconvénient, mais il est difficile de traiter les régions lisses. Par conséquent, la fusion de ces deux approches est proposée pour la détection efficace des modifications de type copier-déplacer [Zheng et al., 2016].

Dans une attaque par épissage, des parties de deux images ou plus sont assemblées pour former une nouvelle image. Pour rendre l'image composée plus réaliste, une opération de post-traitement (mise à l'échelle, recadrage, retouche, rotation, etc.) peut être appliquée à chaque partie de l'image. De plus, après avoir effectué l'opération de collage, une autre opération de post-traitement peut être appliquée pour masquer tout effet perceptible.

Un exemple de cette attaque est donné dans la figure 2.12, où une partie de l'image 2.12a est placée dans l'image 2.12b pour obtenir l'image modifiée 2.12c.

Les techniques basées sur le copier-déplacer ne peuvent pas être utilisées dans le cas de l'épissage, car elles sont basées sur la recherche d'une région correspondante dans l'image, alors que dans le cas de l'épissage, la région altérée est copiée à partir d'une autre image. Par conséquent, la région modifiée a des caractéristiques différentes de celles du reste de l'image.

Les méthodes basées sur l'épissage utilisent diverses caractéristiques telles que la bi-cohérence, la fonction de réponse de la caméra, les coefficients *Discrete Cosine Transform* (DCT) et *Discrete Wavelet Transform* (DWT), les moments invariants de l'image, les descripteurs locaux de Weber, etc. Ng and Chang [2004] utilisent les caractéristiques de magnitude et de phase de la bi-cohérence afin de détecter l'épissage dans les images



(a) (b) (c)

FIGURE 2.12 – Exemple d'image modifiée par épissage (à droite) avec les deux images originales (à gauche et au centre)

numériques. Lin et al. [2005] proposent une méthode pour calculer les fonctions de réponse du dispositif de capture en choisissant différents patchs de l'image. L'incohérence ou l'anomalie de ces zones est ensuite utilisée pour détecter l'épissage. Les méthodes basées sur l'épissage dépendent des caractéristiques intrinsèques des dispositifs plutôt que des caractéristiques réelles des dispositifs de capture.

2.4.3 Discussion

Dans cette sous-section, nous discutons l'utilité des techniques de détection de modifications dans une image dans le contexte des réseaux sociaux. Au vu des spécificités de ces approches, nous pouvons tirer les conclusions suivantes.

Les méthodes actives ne sont pas utiles dans notre contexte, car la stéganographie modifie les valeurs de quelques pixels sans modifier l'aspect visuel de l'image. Si une telle image est diffusée sur les réseaux sociaux, elle ne suscite aucun intérêt car les modifications sont invisibles à l'œil nu. De même que la stéganographie, le tatouage numérique ne change pas l'aspect visuel de l'image. Par conséquent, nous ne nous intéressons pas à ce type de modifications et aux approches permettant de les détecter.

Les méthodes passives basées sur des informations *a priori* sur le format de l'image (par exemple JPEG) ou sur le type d'appareil de capture ne sont pas applicables dans le contexte des réseaux sociaux. En effet, les formats d'image peuvent subir des altérations lors de la publication du message. De même, l'information relative au type d'appareil

de capture de la photo n'est pas non plus toujours disponible. De plus, les plateformes de médias sociaux ont tendance à effacer les métadonnées, en particulier les données *Exchangeable Image File* (Exif). Ces informations générées par l'appareil photo lors de la prise de chaque vue, automatiquement, comme le lieu de la prise ou la date, sont utiles pour la détection de l'altération [12].

Les techniques de détection des images épissées sont difficilement applicables dans le contexte des réseaux sociaux, en raison des transformations (sauvegarde, recadrage) opérées sur ces images par les utilisateurs et des opérations de redimensionnement et de recompression d'images automatiquement appliquées par les plateformes de réseaux sociaux à tous les contenus téléchargés [Zampoglou et al., 2015].

Les techniques existantes de détection de copier-déplacer sont connues pour leur temps de calcul important. Bien qu'elles soient fiables sur des petits jeux de données, elles ne sont pas susceptibles de passer à l'échelle lorsqu'il s'agit d'une grande quantité d'images comme dans le cas des réseaux sociaux [Abd Warif et al., 2016]

De plus, une fausse image dans les réseaux sociaux ne signifie pas forcément une image modifiée. C'est le cas, par exemple, de la publication d'une image relative à un ancien événement pour décrire un événement récent. On parle alors de déformation du contexte d'une image authentique. Par conséquent, une image diffusant de fausses informations dans les médias sociaux peut ne plus contenir de traces détectables d'altération, d'où l'inutilité de toutes ces techniques de détection d'altération dans notre contexte.

En conclusion, pour l'exploitation des données des images, les techniques de détection de modification ne sont pas applicables dans le cas d'images issues des réseaux sociaux. Il est donc quasiment indispensable de se baser sur le contenu de l'image seulement et non sur un type de modification ou un format en particulier.

2.5 Conclusion

Dans ce chapitre, nous avons présenté un état de l'art qui se focalise sur plusieurs aspects. Dans un premier temps, nous avons résumé les approches pratiques de détection de rumeurs proposées par les professionnels et basées principalement sur des techniques

12. http://www.embeddedmetadata.org/social-media-test-procedure.php

de vérification manuelle. Deuxièmement, nous avons proposé et discuté une taxonomie des approches de détection des rumeurs dans les réseaux sociaux en ligne. Ces approches sont très diverses et visent à trouver et utiliser des caractéristiques efficaces des données de la rumeur ou à construire des modèles robustes de détection. Ensuite, nous avons présenté un état de l'art sur les techniques d'analyse de la véracité d'une image numérique. L'étude de l'applicabilité de ces techniques dans le contexte des réseaux sociaux nous montre qu'elles ne sont pas utiles dans le cas des images issues de ces plateformes. En effet, la majorité de ces approches se basent soit sur une connaissance *a priori* ou se concentrent sur un type de modification ou un format particulier d'image.

Cette articulation de l'état de l'art a été choisie car (1) la grande majorité des travaux présentés exploitent le contenu textuel pour la classification des rumeurs. Cependant le contenu visuel, notamment les images ne sont pas suffisamment explorées ; (2) au-delà de l'extraction des caractéristiques visuelles à partir des images, il s'agissait pour nous d'associer la véracité des images dans la détection des rumeurs.

Par ailleurs, les publications issues des réseaux sociaux présentent plusieurs modalités de données. L'analyse des relations entre les données provenant de modalités multiples et le développement de modèles avancés basés sur la fusion pour utiliser ces données constitue, à notre sens, une solution possible pour détecter les rumeurs. C'est ainsi que, dans cette thèse, nous proposons des modèles basés sur l'extraction et la fusion des caractéristiques issues de plusieurs modalités des messages, notamment le texte, le contexte social, les sentiments et le contenu visuel des images.

Cette idée d'apprentissage multimodal constitue l'objet de notre première contribution, MONITOR, qui est présenté dan le chapitre suivant. En plus des caractéristiques textuelles, linguistiques et du contexte social, pour analyser la véracité des images, nous proposons de nouvelles métriques visuelles pour caractériser les images. Ces métriques sont inspirées du domaine de la l'évaluation de la qualité d'image pour être utilisées dans le contexte des réseaux sociaux. Ces métriques évaluent la quantité des dégradations visuelles présentes dans une image et ne nécessitent aucune connaissance *a priori* sur la version originale de l'image. Pour apprendre ces modalités, plusieurs modèles d'apprentissage automatique traditionnel sont utilisés pour la classification des messages.

Chapitre 3

MONITOR : Une plateforme de fusion multimodale basée sur l'apprentissage automatique

Sommaire

Publication(s) associée(s)

— Azri, A., Favre, C., Harbi, N., Darmont, J., Noûs, C. : Monitor : A multimodal fusion framework to assess message veracity in social networks. **ADBIS'2021.** European Conference on Advances in Databases and Information Systems, Tartu, Estonia, 24-26 August 2021, pages 73–87, Springer.

— Azri, A., Favre, C., Harbi, N., Darmont, J. : Including images into message veracity assessment in social media. **INTIS'2019.** International Conference on Innovation and New Trends in Information Technology, December 20-21, 2019, Tangier, Morocco.

Dans le chapitre précédent, nous avons présenté les concepts de base en rapport avec les rumeurs et les média sociaux, ainsi qu'un état de l'art des recherches menées dans le domaine de la détection et de la classification des rumeurs. Dans le présent chapitre, nous abordons la première des contributions de cette thèse, qui porte sur la classification de la véracité des messages de *microblogs*, en utilisant un *framework* de fusion de plusieurs modalités permettant d'apprendre plusieurs algorithmes d'apprentissage automatique.

3.1 Introduction

Comme nous l'avons évoqué précedemment, la plupart des approches de détection automatique des rumeurs abordent cette tâche comme un problème de classification. Elles extraient des caractéristiques de divers aspects des messages, qui sont ensuite utilisées pour entraîner un large éventail de méthodes d'apprentissage automatique [Gupta et al., 2014; Zhou et al., 2019; Horne and Adali, 2017] ou d'apprentissage profond [Wang et al., 2018; Zhou et al., 2020b; Ruchansky et al., 2017]. Les caractéristiques sont généralement extraites du contenu textuel des messages [Pérez-Rosas et al., 2017] et du contexte social [Wu and Liu, 2018]. Cependant, le contenu multimédia des messages, notamment les images qui présentent un ensemble important de caractéristiques, est très peu exploité.

Les images sont très répandues sur les *microblogs* et jouent un rôle crucial dans le processus de la diffusion des nouvelles. À titre d'exemple, dans le jeu de données collecté par Jin et al. [2016b], le nombre moyen de messages avec une image jointe est plus de 11 fois supérieur à celui des messages en texte brut. Dans ce travail, nous appuyons l'hypothèse que l'utilisation des propriétés des images est importante dans la vérification des rumeurs. Avec un contenu visuel riche en informations, les images devraient être utiles pour distinguer les rumeurs. La figure 3.1 montre deux exemples de rumeurs publiées sur Twitter. Dans la figure 3.1(a), il est difficile d'évaluer la véracité du texte, mais l'image probablement manipulée laisse penser à une rumeur. Dans la figure 3.1(b), il est difficile d'évaluer la véracité à partir du texte ou de l'image, car l'image a été détournée de son contexte original.

Proposition et positionnement : Sur la base des observations ci-dessus, nous proposons d'exploiter toutes les modalités des messages de *microblogs* pour vérifier les rumeurs, c'est-à-dire les caractéristiques extraites du contenu textuel et du contexte social

(a) Black clouds in New York City before Sandy!!!

(b) NepalEarthquake 4Years old boy protect his little sister. make me feel so sad

FIGURE 3.1 – Exemples de rumeurs postées sur Twitter

des messages, ainsi que les caractéristiques visuelles et statistiques dérivées des images, très peu utilisées jusqu'à présent. Ensuite, tous les types de caractéristiques doivent être fusionnés pour permettre à un classifieur supervisé d'apprentissage automatique d'évaluer la véracité des messages.

L'originalité de notre approche réside dans le fait que pour les caractéristiques visuelles, nous proposons un ensemble de métriques d'image inspirées du domaine de l'évaluation de la qualité des images (*Image Quality Assessment* – IQA) et nous montrons qu'elles contribuent très efficacement à la vérification de la véracité des messages. Ces métriques estiment le taux de bruit et quantifient la quantité de dégradation visuelle de tout type de modification dans une image. Il est prouvé qu'elles sont de bons indicateurs pour la détection de fausses images, même pour celles générées par des techniques avancées telles que les réseaux adverses génératifs (*Generative Adversarial Networks* – GAN) [Goodfellow et al., 2014].

Nous proposons ainsi le *framework* MONITOR, qui exploite et fusionne tous les types de caractéristiques des messages (c'est-à-dire le texte, le contexte social et les caractéristiques des images) par des algorithmes supervisés d'apprentissage automatique. Ce choix est motivé par le fait que ces techniques permettent d'apporter des éléments d'explicabilité et d'interprétabilité vis-à-vis des décisions prises.

Enfin, nous menons des expériences approfondies sur deux jeux de données du monde réel collectés sur Twitter pour démontrer l'efficacité de notre approche de détection des rumeurs. Nous évaluons les performances de MONITOR par rapport aux méthodes

de référence en apprentissage automatique de la littérature. Les résultats atteints par MONITOR sont meilleurs que ceux de toutes les méthodes de référence, avec un très bon taux de précision de score F_1.

Ce chapitre est organisé comme suit. Dans la section 3.2, nous détaillons MONITOR et plus particulièrement les phases d'extraction et de sélection des caractéristiques. Ensuite, nous présentons et analysons les résultats expérimentaux dans la section 3.3. Enfin, nous concluons ce chapitre et discutons des perspectives dans la section 3.4.

3.2 MONITOR

Les messages de microblogs contiennent de riches ressources multimodales, telles que le contenu du texte, le contexte social entourant le message et l'image jointe. Notre objectif est de tirer parti de ces informations multimodales pour déterminer si un message est vrai ou faux. Sur la base de cette idée, nous proposons une plateforme de vérification de la véracité des messages. La description détaillée de MONITOR est présentée dans cette section.

3.2.1 Aperçu de la fusion multimodale

Nous définissons un message comme un n-uplet de texte, de contexte social et de contenu d'image. MONITOR prend les caractéristiques de ces modalités et vise à apprendre un vecteur de caractéristiques de fusion multimodale comme une agrégation de ces aspects du message. La figure 3.2 présente un aperçu général de MONITOR.

Notre approche comporte deux étapes principales.

1. **Extraction et sélection des caractéristiques.** Nous procédons à l'extraction de plusieurs caractéristiques utiles à partir du texte du message et du contexte social, puis nous exécutons un algorithme de *feature selection* pour identifier les caractéristiques pertinentes, qui forment un premier ensemble de caractéristiques textuelles. Ensuite, à partir de l'image jointe, nous élaborons des statistiques et des caractéristiques visuelles efficaces inspirées du domaine de l'IQA, qui forment un deuxième ensemble de caractéristiques d'images.
2. **Apprentissage du modèle.** Les deux ensembles de caractéristiques textuelles

et images sont ensuite concaténés, standardisés et normalisés pour former le vecteur de fusion comme étant la représentation multimodale finale du message. Ensuite, plusieurs classifieurs d'apprentissage automatique sont utilisés pour apprendre à partir du vecteur de la fusion afin de distinguer la véracité du message (c'est-à-dire vrai ou faux).

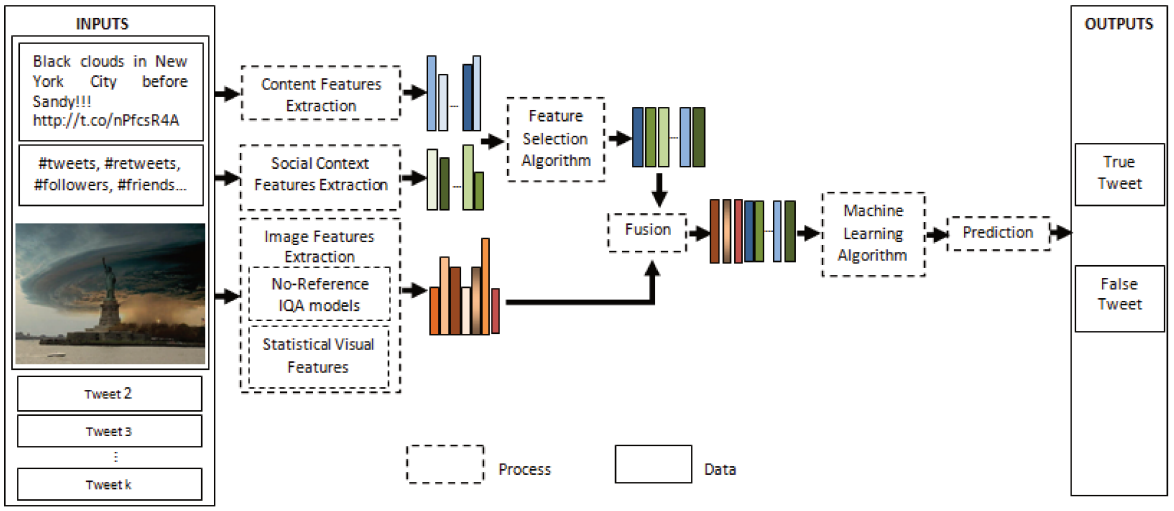


FIGURE 3.2 – Aperçu général du *framework* MONITOR

3.2.2 Extraction et sélection des caractéristiques

L'étape d'extraction de caractéristiques vise à représenter le contenu des messages et les informations auxiliaires connexes dans une structure formelle mesurable. Pour mieux choisir les caractéristiques, nous avons examiné les meilleures pratiques suivies par les professionnels de l'information (par exemple, les journalistes) pour vérifier le contenu généré par les utilisateurs de réseaux sociaux. Nous avons fondé notre réflexion sur les données pertinentes issues d'études journalistiques [Martin, 2014] et du guide de vérification [Silverman, 2014]. Ainsi, nous définissons un ensemble de propriétés qui sont importantes pour extraire les caractéristiques discriminantes des rumeurs. Ces caractéristiques sont principalement dérivées de trois aspects principaux des informations d'actualité : le contenu, le contexte social et le contenu visuel des images.

Le processus de sélection des caractéristiques n'est appliqué qu'aux ensembles de caractéristiques de contenu et de contexte social, vu leur nombre important, afin d'éliminer celles peu pertinentes et qui peuvent avoir un impact négatif sur les performances du *framework*. Parallèlement, puisque nous nous intéressons à l'ensemble des caractéristiques visuelles, nous les conservons toutes dans le processus d'apprentissage.

3.2.2.1 Caractéristiques du contenu du message

Les caractéristiques du contenu sont extraites du texte du message. Visant à susciter beaucoup d'attention et à stimuler l'humeur du public, les textes de rumeurs ont tendance à présenter certaines caractéristiques par rapport aux non-rumeurs. Nous procédons à l'extraction de caractéristiques telles que la longueur du texte d'un *tweet* et le nombre de ses mots. Ces caractéristiques comprennent également des statistiques telles que le nombre de points d'exclamation et d'interrogation, ainsi que des caractéristiques binaires indiquant l'existence ou non d'émoticônes. En outre, d'autres caractéristiques sont extraites de la linguistique du texte, notamment le nombre de mots de sentiment positif et négatif. Pour la langue anglaise, nous utilisons la liste de lexiques d'opinion de Liu et Hu[13], pour l'allemand, les normes affectives de Leipzig [Kanske and Kotz, 2010] et, pour l'espagnol, l'adaptation d'ANEW (*Affective Norms for English Words*) [Redondo et al., 2007]. Des caractéristiques binaires supplémentaires indiquent si le texte contient des pronoms personnels.

La véracité du texte du message peut également être liée à sa lisibilité. Nous calculons un score de lisibilité entre 1 et 100 en utilisant la méthode Flesch Reading Ease [Kincaid et al., 1975]. Ce score est donné par la formule 3.1.

$$score = 203.835 - 1.015 \times \frac{total\ de\ mots}{total\ de\ phrases} - 84.6 \times \frac{total\ syllabes}{total\ de\ mots} \tag{3.1}$$

Plus ce score est élevé, plus le texte est facile à lire. Pour les *tweets* écrits dans une langue pour laquelle les caractéristiques ci-dessus ne peuvent être extraites, nous considérons que les valeurs correspondantes sont manquantes. D'autres caractéristiques sont extraites du contenu informatif fourni par le style de communication spécifique de la plateforme Twitter, comme le nombre de *retweets*, de mentions (@), de *hashtags* (#) et d'URL.

3.2.2.2 Caractéristiques du contexte social

Le contexte social reflète la relation entre les différents utilisateurs et décrit le processus de propagation d'une rumeur. Les caractéristiques du contexte social sont

13. https ://www.cs.uic.edu/ liub/FBS/sentiment-analysis.html#lexicon

donc extraites du comportement des utilisateurs et du réseau de propagation. Nous collectons plusieurs caractéristiques à partir des profils des utilisateurs, telles que le nombre de *followers* et d'amis, le nombre de *tweets* dont l'utilisateur est l'auteur, le nombre de *tweets* qu'il a aimé, si l'utilisateur est vérifié par le média social, et si l'utilisateur a une image de profil.

Nous extrayons également des caractéristiques de l'arbre de propagation qui peut être construit à partir des *tweets* et des *retweets* d'un message, comme la profondeur de l'arbre des *retweets*. Les tableaux 3.1 et 3.2 décrivent un ensemble de caractéristiques de contenu et de contexte social extraites de chaque message.

TABLE 3.1 – Caractéristiques du contenu

Description
of characters, words
of question mark (?), exclamation mark (!)
of uppercase characters in the tweet text
of positive, negative sentiment words
of mentions(@username), hashtags(#link), URLs
of happy, sad mood emoticon
first, second, third order pronoun
The readability score of the tweet text

TABLE 3.2 – Caractéristiques du contexte social

Description
of followers, friends, posts the user has
Friends/followers ratio, times listed the user has
of re-tweets, likes that the tweet has obtained
Whether the user shares a homepage URL
Whether the user has their own profile image
Whether the author has a verified account
of Tweets the user has liked

Afin d'améliorer les performances de MONITOR, nous exécutons un algorithme de sélection de caractéristiques sur les ensembles de caractéristiques listés dans les tableaux 3.1 et 3.2. Les détails du processus de sélection des caractéristiques sont abordés dans la section 3.3.

3.2.2.3 Caractéristiques des images

Pour différencier les vraies et les fausses images dans les messages, nous proposons d'exploiter les caractéristiques du contenu visuel et les statistiques visuelles qui sont extraites des images jointes.

Caractéristiques du contenu visuel : Habituellement, un utilisateur de *microblog* décide de la véracité d'une image sur la base de sa perception subjective, cela nous amène à poser la question suivante : comment représenter quantitativement la perception humaine de la qualité d'une image ? La qualité d'une image désigne la quantité de dégradations visuelles de tous types présentes dans une image, telles que le bruit, le flou, la décoloration, etc.

Le domaine de l'IQA vise à quantifier la perception humaine de la qualité de l'image en fournissant un score objectif des dégradations de l'image basé sur des modèles de calcul [Maître, 2017]. Ces dégradations sont introduites lors de différentes étapes de traitement, telles que l'acquisition, le traitement, la compression, le stockage, la transmission, la décompression, l'affichage ou même l'impression des images. Inspirés par la pertinence potentielle des métriques IQA dans notre contexte, nous les utilisons de manière originale pour un objectif différent de celui pour lequel elles ont été créées. Plus précisément, nous pensons que l'évaluation quantitative de la qualité d'une image peut être utile pour la détection de sa véracité.

L'IQA se divise principalement en deux types de techniques. (1) Dans les algorithmes d'évaluation avec référence (*full-reference evaluation*), la métrique est calculée en comparant l'image d'entrée avec une image de référence sans aucune distorsion, par exemple, comparer une image de référence avec sa version compressée par JPEG. (2) Dans les algorithmes d'évaluation sans référence (*no-reference evaluation*), l'entrée est la seule image dont on veut mesurer la qualité (sans rien connaître de l'image de référence). Ces algorithmes comparent les caractéristiques statistiques de l'image d'entrée à un ensemble de caractéristiques dérivées d'une base de données d'images.

Dans notre cas, nous ne disposons pas de la version originale de l'image postée par l'utilisateur sur le *microblog*. Par conséquent, l'approche qui convient à notre contexte doit utiliser les métriques d'évaluation sans référence. À cette fin, nous utilisons trois algorithmes de cette famille, dont l'efficacité a été démontrée dans des travaux de

recherche de l'IQA.

— Le *Blind/Referenceless Image Spatial Quality Evaluator* (BRISQUE) [Mittal et al., 2011] est entraîné sur une base de données d'images avec des distorsions connues, et est limité à l'évaluation de la qualité des images avec le même type de distorsion. BRISQUE est *opinion-aware*, ce qui signifie que des scores de qualité subjectifs données par des experts sont associés aux images d'entraînement.
— Le *Naturalness Image Quality Evaluator* (NIQE) [Mittal et al., 2012] est entraîné sur une base de données d'images sans aucune distorsion. Il peut mesurer la qualité d'images présentant une distorsion arbitraire. NIQE est une métrique *opinion-unaware*, c'est-à-dire qu'elle ne tient pas compte des scores de qualité subjectifs des experts.
— Le *Perception based Image Quality Evaluator* (PIQE) [Venkatanath et al., 2015] est un algorithme non-supervisé (c'est-à-dire qu'il ne nécessite pas de modèle entraîné) et ne tient pas compte des scores de qualité subjectifs des experts humains. PIQE peut mesurer la qualité d'images présentant une distorsion arbitraire.

À titre d'exemple, la figure 3.3 illustre le score BRISQUE calculé pour une image originale et ses versions déformées (déformations par compression, bruit et flou). Le score BRISQUE est un scalaire non négatif compris dans l'intervalle [1, 100]. Des valeurs plus faibles du score reflètent une meilleure qualité perceptuelle de l'image. Nous pouvons constater que, par rapport à l'image originale, la qualité perceptuelle se dégrade pour chaque type d'altération.

Image originale	Compression JPEG	Bruit gaussien	Flou médian
13.7215	22.6603	28.5840	41.5620

FIGURE 3.3 – Score BRISQUE calculé pour une image originale et ses versions altérées

Les métriques IQA de sans référence sont également de bons indicateurs pour d'autres types de modifications d'images, comme les images générées par des réseaux adverses génératifs (GAN). Ces techniques permettent de modifier le contexte et la sémantique

des images de manière très réaliste. Contrairement à de nombreuses tâches d'analyse d'images, où les images de référence et les images reconstruites sont disponibles, les images générées par les GAN peuvent ne pas avoir d'image de référence. C'est la raison principale de l'utilisation des techniques d'IQA sans référence pour évaluer ce type de fausses images. La figure 3.4 montre le score BRISQUE calculé pour des images réelles et fausses générées par la translation d'image à image basée sur des GAN (convertir une image d'un domaine source X en un domaine cible Y en l'absence d'exemples appariés) [Zhu et al., 2017].

Image réelle : 17.7778 | Image fausse : 22.0260 | Image réelle : 12.5000 | Image fausse : 22.5279

FIGURE 3.4 – Score BRISQUE calculé pour de vraies et fausses images GAN

Caractéristiques statistiques du contenu visuel : À partir des images jointes, nous définissons quatre caractéristiques statistiques à partir de deux aspects.

— *Nombre d'images :* Un utilisateur peut poster une, plusieurs ou aucune image. Pour désigner cette caractéristique, nous comptons le nombre total d'images dans une rumeur et le ratio de messages contenant plus d'une image.
— *Diffusion d'images :* Pendant un événement, certaines images sont très répandues et génèrent plus de commentaires que d'autres. Le ratio de ces images est calculé pour indiquer cette caractéristique.

Le tableau 3.3 illustre la description de nos caractéristiques visuelles et statistiques. Nous utilisons toutes ces caractéristiques d'images dans le processus d'apprentissage.

3.2.3 Apprentissage du modèle

Jusqu'à présent, nous avons obtenu un premier ensemble de caractéristiques textuelles pertinentes grâce à un processus de sélection des caractéristiques. Nous dis-

TABLE 3.3 – Description des caractéristiques des images

Type	Feature	Description
Visual Features	BRISQUE	The BRISQUE score of a given image
	PIQE	The PIQE score of a given image
	NIQE	The NIQE score of a given image
Statistical Features	Count_Img	The number of all images in a news event
	Ratio_Img1	The ratio of the multi-image tweets in all tweets
	Ratio_Img2	The ratio of image number to tweet number
	Ratio_Img3	The ratio of the most widespread image in all distinct images

posons également d'un deuxième ensemble de caractéristiques d'image composé de caractéristiques statistiques et visuelles. Ces deux ensembles de caractéristiques sont normalisés et concaténés pour former la représentation multimodale d'un message donné, qui est apprise par un classifieur supervisé. Plusieurs algorithmes d'apprentissage automatique peuvent être mis en œuvre pour la classification de la véracité des messages. Nous examinons les algorithmes qui offrent les meilleures performances dans la section 3.3.

3.3 Expérimentation

Pour démontrer l'efficacité de notre modèle, nous menons dans cette section des expériences approfondies sur deux jeux de données publics du monde réel. Tout d'abord, nous présentons les statistiques relatives aux jeux de données que nous utilisons. Ensuite, nous décrivons les paramètres expérimentaux. Nous passons brièvement en revue les caractéristiques de référence pour la vérification des nouvelles et nous sélectionnons les meilleures de ces caractéristiques textuelles comme *baselines*. Enfin, nous présentons les résultats expérimentaux et analysons les caractéristiques pour obtenir des éléments de compréhension avec MONITOR.

3.3.1 Jeux de données

Pour évaluer les performances de MONITOR, nous menons des expériences sur deux jeux de données publics de détection des rumeurs. Les statistiques détaillées de ces deux jeux de données sont listées dans le tableau 3.4.

TABLE 3.4 – Statistiques des jeux de données MediaEval et FakeNewsNet

Dataset	Set	Tweets		Images
		Real	Fake	
MediaEval	Ensemble d'entraînement	5,008	6,841	361
	Ensemble de test	1,217	717	50
FakeNewsNet	Ensemble d'entraînement	25,673	19,422	47,870
	Ensemble de test	6,466	4,808	11,968

MediaEval [14] [**Boididou et al., 2015**] est collecté à partir de Twitter et comprend les trois caractéristiques : texte, contexte social et images. Il est conçu pour la vérification de la véracité des messages. Le jeu de données comprend deux parties : un jeu de développement contenant environ 9000 faux *tweets* et 6000 vrais *tweets* provenant de 17 évènements ; un jeu de test contenant environ 2000 *tweets* provenant d'un autre lot de 35 évènements liés aux rumeurs. Nous supprimons les *tweets* sans texte ni image, obtenant ainsi un jeu de données final comprenant 411 images distinctes associées à 6225 vrais et 7558 faux *tweets*, respectivement.

FakeNewsNet [15] [**Shu et al., 2020**] est l'un des principaux référentiels de détection des fausses nouvelles. Les articles de fausses et de vraies nouvelles sont collectés sur les sites web de vérification des faits PolitiFact [16] et GossipCop [17]. Comme nous sommes particulièrement intéressés par les images dans ce travail, nous procédons à l'extraction et à l'exploitation des informations relatives aux images de tous les *tweets*. Étant donné que la plupart des algorithmes de classification de l'apprentissage automatique sont conçus en partant de l'hypothèse d'une distribution équilibrée des classes, nous

14. https://github.com/MKLab-ITI/image-verification-corpus/tree/master/mediaeval2015
15. https://github.com/KaiDMML/FakeNewsNet
16. https://www.politifact.com/
17. https://www.gossipcop.com/

choisissons au hasard 2566 vraies nouvelles et 2587 fausses. Après avoir supprimé les *tweets* sans image, nous obtenons 56 369 *tweets* et 59 838 images.

3.3.2 Paramètres expérimentaux

3.3.2.1 Mesures de performance

Nous rapportons la performance quantitative de chaque modèle en termes de précision (Accuracy), de précision (pecision), de rappel et de score F_1, que nous détaillons par la suite. Pour toutes ces mesures, nous utilisons la macro-moyenne des deux classes.

Étant donné que notre objectif est la détection des fausses information, notre « classe positive » est constituée des faux messages (définition 2). Par conséquent, un vrai positif (*True Positive* – TP) signifie que le message est réellement faux et également prédit comme faux, tandis que le faux positif (False Positive – FP) indique que le message est réellement vrai, mais prédit comme faux. Le vrai négatif (*True Negative* – TN) et le faux négatif (*False Negative* – FN) sont définis en conséquence. La précision (*accuracy* – ACC) est le nombre d'instances correctement prédites sur l'ensemble des instances (équation 3.2).

$$ACC = \frac{TP + TN}{TP + FN + TN + FP} \tag{3.2}$$

La précision (*precision*) est le rapport entre le nombre d'instances correctement prédites et toutes les instances prédites pour une classe donnée. Pour les classes vraies et fausses, nous présentons cette métrique comme $Prec(T)$ et $Prec(F)$, respectivement. Par conséquent, la macro-moyenne de la précision $Prec$ est la moyenne de $Prec(T)$ et de $Prec(F)$ (équation 3.3).

$$Prec(F) = \frac{TP}{TP + FP},\ Prec(V) = \frac{TN}{TN + FN},\ Prec = \frac{Prec(F) + Prec(V)}{2} \tag{3.3}$$

Le rappel (*recall*) représente le rapport entre le nombre d'instances prédites correctement et de toutes les instances appartenant à une classe donnée. Pour les classes fausses

et vraies, nous présentons cette métrique respectivement comme $Rec(F)$ et $Rec(V)$. Par conséquent, la macro-moyenne du rappel Rec est la moyenne de $Rec(F)$ et $Rec(V)$ (équation 3.4).

$$Rec(F) = \frac{TP}{TP + FN}, \; Rec(V) = \frac{TN}{TN + FP}, \; Rec = \frac{Rec(F) + Rec(V)}{2} \tag{3.4}$$

Enfin, le score F_1 est la moyenne harmonique de la précision et du rappel (équation 3.5).

$$F_1 = \frac{2 \times Prec \times Rec}{Prec + Rec} \tag{3.5}$$

3.3.2.2 Caractéristiques de référence

Nous comparons l'efficacité de notre collection de caractéristiques avec les meilleures caractéristiques textuelles de la littérature. Tout d'abord, nous adoptons les 15 meilleures caractéristiques extraites par Castillo et al. [2011] pour analyser la crédibilité des nouvelles diffusées sur Twitter, qui sont considérées comme un standard dans ce domaine. Nous collectons également un total de 40 caractéristiques textuelles supplémentaires issues de plusieurs travaux de la littérature [Gupta et al., 2013, 2012; Kwon et al., 2013; Wu et al., 2015] et qui sont extraites du contenu textuel, des informations des utilisateurs et des propriétés du réseau de propagation (tableau 5.3).

3.3.2.3 Ensembles de caractéristiques

Les caractéristiques intitulées **Textual** sont les meilleures caractéristiques sélectionnées parmi celles du contenu du message et du contexte social (tableaux 3.1 et 3.2). Nous les sélectionnons à l'aide de la méthode du ratio de gain d'information (*information gain ratio*) [Karegowda et al., 2010], qui permet de sélectionner un sous-ensemble de 15 caractéristiques textuelles pertinentes dont le gain d'information est supérieur à zéro (tableau 3.6). Les caractéristiques intitulées **Image** sont toutes les caractéristiques d'image listées dans le tableau 3.3. Les caractéristiques intitulées **MONITOR** sont l'ensemble de caractéristiques que nous proposons, consistant en la fusion des ensembles de caractéristiques textuelles et d'images. Les caractéristiques intitulées **Castillo** sont

TABLE 3.5 – Caractéristiques tirées de la littérature

Feature
Fraction of (?), (!) Mark, # of messages
Average # of words, char lengths
Fraction of 1st, 2nd, 3rd pronouns
Fraction of URLs, @, #
Count of distinct URLs, @, #
Fraction of popular URLs, @, #
The tweet includes pictures
Average sentiment score
Fraction of positive and negative tweets
of distinct people, loc, org
Fraction of people, loc, org
Fraction of popular people, loc, org
of Users, fraction of popular users
of followers, followees, posted tweets
The user has a Facebook link
Fraction of verified users, org
of comments on the original message
Time between original message and repost

les 15 meilleures caractéristiques textuelles extraites par Castillo et al. [2011]. Enfin, les caractéristiques intitulées **Wu** sont les 40 caractéristiques textuelles identifiées dans la littérature.

3.3.2.4 Construction du modèle

Nous ne pouvons pas savoir à l'avance quel algorithme sera adéquat pour notre problème, ni quelle configuration utiliser. En analysant les deux jeux de données, nous avons constaté que les classes sont partiellement linéairement séparables dans certaines dimensions. Ainsi, nous imlémentons et évaluons un mélange d'algorithmes linéaires et non linéaires simples en utilisant la bibliothèque Python scikit-learn [Pedregosa et al., 2011]. Les meilleurs résultats sont obtenus par quatre algorithmes de classification supervisée : *Classification and Regression Trees* (CART), *k-Nearest Neighbors* (KNN), *Support Vector Machines* (SVM) et *Random Forest* (RF) (Tableau 3.8).

TABLE 3.6 – Meilleures caractéristiques textuelles sélectionnées

MediaEval	**FakeNewsNet**
Tweet_Length	Tweet_Length
Num_Negwords	Num_Words
Num_Mentions	Num_Questmark
Num_URLs	Num_Upperchars
Num_Words	Num_Exclmark
Num_Upperchars	Num_Hashtags
Num_Hashtags	Num_Negwords
Num_Exclmark	Num_Poswords
Num_Thirdpron	Num_Followers
Times_Listed	Num_Friends
Num_Tweets	Num_Favorites
Num_Friends	Times_Listed
Num_Retweets	Num_Likes
Has_Url	Num_Retweets
Num_Followers	Num_Tweets

Il y a lieu de préciser que, pour obtenir les meilleurs résultats de chaque modèle, nous avons procédé à l'optimisation des hyperparamètres de chaque modèle (tableau 3.7) en testant plusieurs configurations possibles à l'aide de la fonction GridSearchCV de scikit-learn. En conséquence, nous réalisons l'entraînement et la validation pour chaque modèle par une validation croisée (*5-fold cross-validation*) afin d'obtenir des résultats stables. Notons que, pour MediaEval, nous conservons le même schéma de répartition des données. Pour FakeNewsNet, nous divisons les données en sous-ensembles d'entraînement et de test avec le ratio 0,8/0,2. Le tableau 3.8 présente les résultats de nos expériences.

3.3.3 Résultats de la classification

À partir des résultats de classification consignés dans le tableau 3.8, nous pouvons faire les observations suivantes.

TABLE 3.7 – Espace de configuration des hyperparamètres

Model	Main hyper-parameters	Type	Search space
CART	max_depth	Discrete	[1,21]
	criterion	Categorical	['gini','entropy']
KNN	n_neighbors	Discrete	[1,21]
SVM	C	Discrete	[0.1,2.0]
	γ (RBF kernel)	Discrete	[0.1,1.0]
	Kernel	Categorical	['linear', 'poly', 'rbf','sigmoid']
RF	n_estimators	Discrete	[10,500]
	max_depth	Discrete	[3,20]

TABLE 3.8 – Performances des différents modèles d'apprentissage automatique

Model	**Features**	**MediaEval**				**FakeNewsNet**			
		Acc	**Prec**	**Rec**	$\boldsymbol{F_1}$	**Acc**	**Prec**	**Rec**	$\boldsymbol{F_1}$
CART	Textual	0.673	0.672	0.771	0.718	0.699	**0.647**	0.652	0.65
	Image	0.632	0.701	0.639	0.668	0.647	0.595	0.533	0.563
	MONITOR	**0.746**	**0.715**	**0.897**	**0.796**	**0.704**	0.623	**0.716**	**0.667**
	Castillo	0.643	0.711	0.648	0.678	0.683	0.674	0.491	0.569
	Wu	0.65	0.709	0.715	0.711	0.694	0.663	0.593	0.627
KNN	Textual	0.707	0.704	0.777	0.739	0.698	0.67	0.599	0.633
	Image	0.608	0.607	0.734	0.665	0.647	0.595	0.533	0.563
	MONITOR	**0.791**	**0.792**	**0.843**	**0.817**	**0.758**	**0.734**	**0.746**	**0.740**
	Castillo	0.652	0.698	0.665	0.681	0.681	0.651	0.566	0.606
	Wu	0.668	0.71	0.678	0.693	0.694	0.663	0.593	0.627
SVM	Textual	0.74	0.729	0.834	0.779	0.658	0.657	0.44	0.528
	Image	0.693	0.69	0.775	0.73	0.595	0.618	0.125	0.208
	MONITOR	**0.794**	**0.767**	**0.881**	**0.82**	**0.771**	**0.743**	**0.742**	**0.743**
	Castillo	0.702	0.761	0.716	0.737	0.629	0.687	0.259	0.377
	Wu	0.725	0.763	0.73	0.746	0.642	0.625	0.394	0.484
RF	Textual	0.747	0.717	0.879	0.789	0.778	0.726	0.768	0.747
	Image	0.652	0.646	0.771	0.703	0.652	0.646	0.771	0.703
	MONITOR	**0.962**	**0.965**	**0.966**	**0.965**	**0.889**	**0.914**	**0.864**	**0.889**
	Castillo	0.702	0.727	0.723	0.725	0.714	0.669	0.67	0.67
	Wu	0.728	0.752	0.748	0.75	0.736	0.699	0.682	0.691

3.3.3.1 Comparaison des performances

Avec MONITOR, l'utilisation des caractéristiques textuelles et visuelles permet à tous les modèles de classification d'obtenir de meilleures performances (selon les quatre mesures de comparaison) comparativement à tous les modèles de référence. Parmi les quatre modèles de classification, RF génère la meilleure précision (*accuracy*) : 96,2 % sur MediaEval et 88,9 % sur FakeNewsNet, soit 26 % et 18 % de mieux que Castillo et 24 % et 15 % que Wu, toujours sur MediaEval et FakeNewsNet, respectivement.

Par rapport à l'ensemble des 15 meilleures caractéristiques textuelles, RF améliore la précision (*accuracy*) de plus de 22 % et de 10 % avec les caractéristiques d'image uniquement. De même, les trois autres modèles obtiennent un gain de précision (*accuracy*) compris entre 5 et 9 % sur MediaEval et entre 5 et 6 % sur FakeNewsNet. Finalement, tous les algorithmes de classification dégradent la précision (*accuracy*) lorsqu'ils utilisent uniquement des caractéristiques d'image.

Ainsi, si les caractéristiques des images jouent un rôle crucial dans la vérification des rumeurs, nous ne devons pas ignorer l'efficacité des caractéristiques textuelles. Les rôles des caractéristiques d'image et des caractéristiques textuelles sont en effet complémentaires. Lorsque les deux ensembles de caractéristiques sont combinés (résultats de MONITOR), les performances sont considérablement améliorées.

3.3.3.2 Illustration par l'exemple

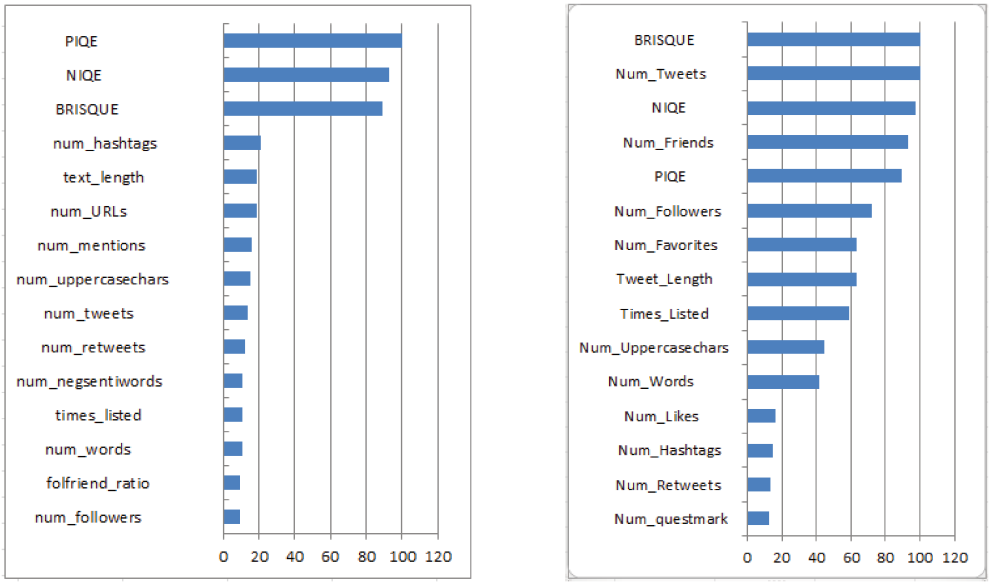
Pour montrer plus clairement la complémentarité entre le texte et les images, nous comparons les résultats obtenus avec MONITOR et les approches monomodales (texte seul ou image seule). Les deux messages de rumeurs de la figure 3.1 (section 4.1) sont correctement détectés comme faux par MONITOR, alors que l'utilisation des modalités textuelles ou images uniquement donne un résultat vrai.

Dans le *tweet* de la figure 3.1(a) (page 52), le contenu textuel décrit uniquement l'image jointe sans donner aucun signe sur la véracité du *tweet*. C'est pourquoi la modalité textuelle identifie ce *tweet* comme réel. C'est l'image jointe qui semble plutôt suspecte et pourrait très probablement être une fausse image. En combinant les contenus textuel et image, MONITOR peut identifier la véracité du *tweet* avec un score élevé, en exploitant certains indices de l'image pour obtenir la bonne classification.

Le *tweet* de la figure 3.1(b) (page 52) est un exemple de rumeur correctement classée par MONITOR comme étant fausse, mais incorrectement classée en utilisant uniquement la modalité visuelle. L'image semble normale, mais sa sémantique complexe est très difficile à saisir par la modalité image. Cependant, les mots traduisant de fortes émotions dans le texte indiquent qu'il pourrait s'agir d'un message suspect. En combinant les modalités texte et image, MONITOR peut classer le *tweet* avec un score de précision élevée. Ce *tweet* présente un type particulier de rumeur qui est très difficile à identifier, car l'image jointe a été détournée de son contexte original (les deux enfants ont été photographiés au Vietnam en 2007) et utilisé dans un autre contexte (le tremblement de terre au Népal en 2015).

3.3.4 Analyse des caractéristiques

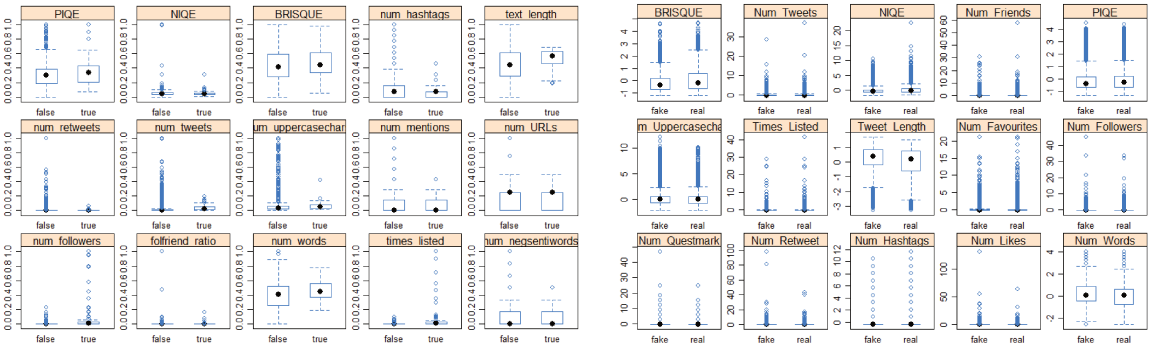
L'avantage de notre approche est que nous pouvons obtenir certains éléments d'interprétabilité. À cette fin, nous effectuons une analyse pour illustrer l'importance de chaque ensemble de caractéristiques. Nous décrivons les 15 premières caractéristiques les plus importantes obtenues par le modèle RF dans la figure 5.3, qui montre que, pour les deux jeux de données MediaEval et FakeNewsNet, les caractéristiques visuelles figurent parmi les cinq premières caractéristiques. Les autres caractéristiques sont un mélange de caractéristiques de contenu textuel et de contexte social.



(a) MediaEval (b) FakeNewsNet

FIGURE 3.5 – Caractéristiques importantes de l'algorithme RF

Enfin, pour illustrer la capacité de discrimination de ces caractéristiques, nous avons déployé des diagrammes en boîtes pour chacune des 15 premières variables sur les deux jeux de données. La figure 3.6 montre que plusieurs caractéristiques présentent une différence significative entre les classes fausse et vraie, ce qui explique les résultats obtenus.



(a) MediaEval (b) FakeNewsNet

FIGURE 3.6 – Distribution des classes vraie et fausse pour les 15 caractéristiques les plus importantes

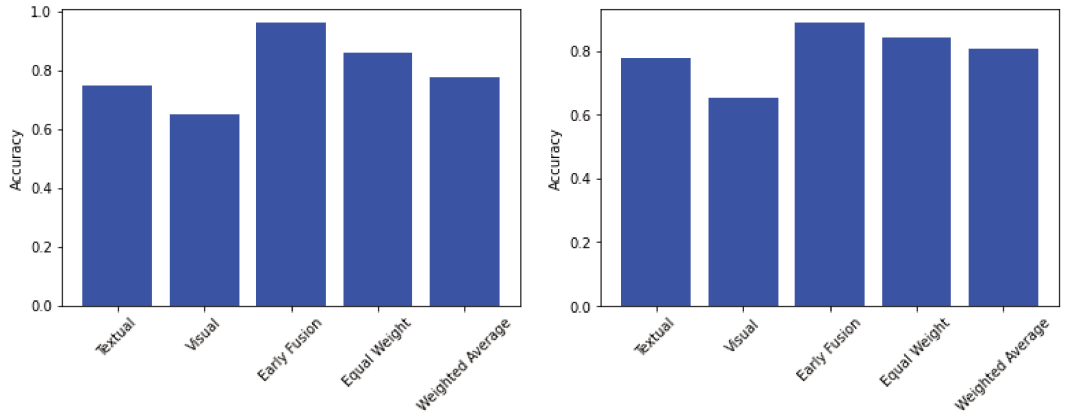
3.3.5 Fusion précoce et tardive

Dans nos expériences précédentes, nous avons simplement fusionné les caractéristiques visuelles et textuelles en un seul vecteur selon la méthode dite de fusion précoce (*early fusion*). La fusion tardive (*late fusion*) est une autre stratégie conventionnelle permettant de combiner les capacités représentatives de diverses caractéristiques. Contrairement à la fusion précoce, la fusion tardive est basée sur la combinaison au niveau des classifieurs.

Dans cette partie des expériences, nous entraînons deux algorithmes RF, le premier pour les caractéristiques textuelles et le second pour les caractéristiques visuelles. Les probabilités de prédiction des deux RF sont ensuite combinées avec une pondération égale (*equal weights*) ou optimisées (*weighted average*) au moyen d'un modèle de régression logistique pour produire les résultats finaux.

À partir de la figure 3.7, nous pouvons voir que, pour les deux jeux de données, la stratégie de fusion précoce et les deux stratégies de fusion tardive, c'est-à-dire la pondération égale et la pondération optimisée, améliorent la prédiction avec des taux différents en utilisant séparément deux ensembles de caractéristiques. De plus, la fusion

précoce donne les meilleurs résultats en termes de précision (*accuracy*) comparativement aux autres, tandis que pour les deux techniques de fusion tardive, la pondération égale est légèrement plus efficace que la pondération optimisée.



(a) MediaEval (b) FakeNewsNet

FIGURE 3.7 – Performances des stratégies de la fusion précoce et tardive avec MONITOR

Les performances de la fusion tardive sont inférieures à celles de la fusion précoce car, lorsque nous formons deux modèles séparément sur les caractéristiques visuelles et textuelles, une certaine information de dépendance entre les caractéristiques est perdue. En pratique, il existe une certaine corrélation entre les deux ensembles de caractéristiques, par exemple entre BRISQUE et Num_Mention ou entre PIQE et Text_Length.

3.4 Conclusion

Dans ce chapitre, nous avons présenté notre première contribution pour la classification de la véracité des messages de *microblogs*. Il s'agit d'une plateforme de fusion de plusieurs modalités appelée MONITOR qui utilise des caractéristiques extraites du contenu textuel des messages, du contexte social et des caractéristiques de l'image. Plusieurs modèles, avec des hyperparamètres optimisés, de l'apprentissage automatique ont été utilisés pour fusionner ces caractéristiques et prédire la véracité des messages. Concrètement, les caractéristiques des images constituent l'essence de cette contribution, étant donné que l'importance et l'utilité des caractéristiques du texte et du contexte social pour la classification des rumeurs ont été déjà démontrées par la majorité des

travaux de recherche.

En effet, les caractéristiques visuelles retenues sont inspirées du domaine de l'évaluation de la qualité d'image pour être utilisées dans le contexte de la vérification de la véracité de messages des *microblogs*. Ces métriques, qui estiment la quantité de dégradation visuelle dans une image, constituent de bons indicateurs pour la détection de la véracité des images. Les expériences approfondies menées sur deux jeux de données ont démontré la pertinence de notre proposition, avec les conclusions suivantes. (1) Les caractéristiques de l'image que nous introduisons jouent un rôle clé dans l'évaluation de la véracité des messages. (2) Aucun ensemble homogène de caractéristiques ne peut générer les meilleurs résultats à lui seul. Les rôles des caractéristiques d'images et des caractéristiques textuelles sont en effet complémentaires. (3) Avec l'utilisation de toutes les familles de caractéristiques, MONITOR surpasse les méthodes d'apprentissage automatique de référence avec une précision de classification supérieure à 96 % sur MediaEval et 89 % sur FakeNewsNet.

Par ailleurs, nous avons comparé notre stratégie précoce de fusion des modalités avec une autre stratégie conventionnelle de fusion, à savoir la fusion tardive basée sur la fusion des classifieurs. Les résultats ont démontré que la fusion précoce est meilleure que les deux variantes de la fusion tardive, à savoir la pondération égale et la pondération optimisée.

Ce premier travail concernant la classification des rumeurs dans les *microblogs* ouvrent diverses perspectives de travail. Nous identifions deux pistes pour l'amélioration des performances de MONITOR. Dans cette contribution, nous avons utilisé des algorithmes avec optimisation des hyperparamètres pour la fusion et la classification des ensembles de caractéristiques extraites et pour choisir le modèles le plus performant. Néanmoins, cela peut impliquer beaucoup d'essais et d'erreurs pour trouver un modèle qui fonctionne suffisamment bien. Une autre approche possible est de combiner les prédictions de différents modèles en utilisant des techniques d'apprentissage ensembliste. L'idée est que chaque modèle a ses propres performances qui le diffèrent des autres. Autrement dit, les erreurs de prédictions faites par les modèles sont non-corrélées ou ils ont une faible corrélation, ce qui devrait améliorer les performances de MONITOR. Cette perspective sera abordée dans le chapitre suivant par la proposition et l'évaluation de plusieurs algorithmes d'apprentissage ensembliste.

La deuxième piste de recherche est d'explorer les techniques d'apprentissage profond, qui se révèlent capables d'apprendre des représentations de textes et d'images d'une

manière plus exacte que celles développées par l'ingénierie des caractéristiques. Plus précisément, les réseaux de neurones convolutifs (CNN) ont démontré leurs capacités en matière de représentation d'images [Simonyan and Zisserman, 2014], tandis que les réseaux de neurones récurrents (RNN) sont largement employés récemment dans la représentation du texte [Bahdanau et al., 2014]. Cette perspective, qui rentre toujours dans l'idée d'apprentissage multimodal, sera présentée dans le chapitre 5 par la proposition d'un modèle neuronal profond appelé deepMONITOR. Ce modèle est capable de fusionner et d'apprendre efficacement des représentations issues de plusieurs modalités, notamment le texte, les images et les signaux sentimentaux.

Chapitre 4

Apprentissage ensembliste pour améliorer les performances de MONITOR

Sommaire

Publication(s) associée(s)

Azri, A., Favre, C., Harbi, N., Darmont, J., Noûs, C. : Rumor Classification by Multimodal Fusion and Ensemble Learning. (soumis en Décembre 2021 au journal Information Systems Frontiers, accepté avec une révision mineure en Avril 2022).

Dans le chapitre précédent, nous avons présenté MONITOR, un *framework* de fusion multimodale pour la classification de la véracité des rumeurs dans les réseaux sociaux. Dans ce chapitre, nous décrivons la deuxième contribution de cette thèse, qui porte sur l'utilisation de techniques d'apprentissage ensembliste pour améliorer les performances de MONITOR.

4.1 Introduction

MONITOR utilise des algorithmes classiques d'apprentissage automatique pour la classification de la véracité des messages. Un problème majeur des modèles d'apprentissage automatique est que des modèles différents ont des performances différentes sur le même corpus de données. Leurs performances sont sensibles aux propriétés du corpus, comme la taille du corpus et la distribution des instances dans les classes. Par exemple, la performance de KNN dépend du nombre k de points les plus proches dans le jeu de données. Les SVM souffrent d'instabilité numérique lors de la résolution des problèmes d'optimisation [Pham et al., 2021]. Autrement dit, certains de ces algorithmes sont particulièrement performants sur un jeu de données, mais ils peuvent donner des performances moyennes voire médiocres sur d'autres jeux de données.

Un processus d'apprentissage implique souvent le choix d'un modèle, peut être le meilleur que nous puissions trouver, parmi plusieurs autres. Cette manière de faire, demande beaucoup de temps et d'efforts pour optimiser un nombre important d'hyperparamètres des algorithmes candidats. Une autre alternative, préférable, consiste à utiliser un groupe de classifieurs et de prendre une décision collective ou d'équipe, plutôt que de baser une décision sur un classifieur individuel.

L'apprentissage ensembliste est considéré comme une technique efficace, qui peut améliorer la performance des modèles individuels d'apprentissage automatique, également appelés modèles de base ou prédicteurs de base, en regroupant les prédictions de ces modèles d'une certaine manière. Il vise à exploiter la diversité des prédicteurs de base pour traiter plusieurs types d'erreurs afin d'augmenter la performance globale.

Bien que les méthodes d'apprentissage ensemblistes aient montré de bonnes performances dans diverses applications, très peu d'efforts ont été consentis par les travaux de la littérature (section 4.2) pour explorer l'utilité de ces modèles d'apprentissage non conventionnels pour la détection des rumeurs.

Proposition et positionnement. Les travaux existants qui utilisent l'apprentissage ensembliste pour la classification de rumeurs reposent principalement sur des modèles de base homogènes, comme les techniques de *boosting* et de *bagging*, qui sont présentés dans la section suivante. Contrairement à ces approches, nous utilisons des modèles de base hétérogènes en nous basant sur les techniques de *stacking*. L'avantage de notre approche est qu'elle permet d'exploiter les capacités de plusieurs modèles sur une tâche de classification et que le résultat de la prédiction est meilleur que n'importe quel modèle individuel.

C'est ainsi que, dans ce chapitre, nous utilisons les approches d'apprentissage ensembliste pour améliorer les performances de MONITOR. Pour ce faire, nous développons cinq algorithmes de *metalearning* : *soft voting*, *weighted average voting*, *stacking*, *blending* et un *super learner ensemble*. Ces modèles ensemblistes utilisent les quatre algorithmes d'apprentissage automatique (KNN, CART, SVM et RF) utilisés par MONITOR comme modèles de base.

Pour éviter toute ambiguïté, le terme *metalearning* dans l'apprentissage automatique fait référence aux algorithmes d'apprentissage qui apprennent à partir d'autres algorithmes d'apprentissage. Le plus souvent, cela signifie l'utilisation d'algorithmes d'apprentissage automatique qui apprennent à combiner au mieux les prédictions d'autres algorithmes, dans le domaine de l'apprentissage ensembliste.

Nous menons des expériences poussées sur deux jeux de données du monde réel collectés sur Twitter. Nous montrons l'utilité de l'apprentissage ensembliste pour la tâche de classification de rumeurs en comparant les performances des cinq algorithmes de *metalearning* avec le meilleur modèle individuel obtenu avec MONITOR. Les résultats obtenus montrent clairement que l'approche ensembliste peut améliorer notablement les performances de MONITOR.

Ce chapitre est organisé de la manière suivante. Dans la section 4.2, nous présentons une synthèse de l'état de l'art de l'apprentissage ensembliste. Dans la section 4.3, nous décrivons les algorithmes de l'approche que nous proposons. Nous présentons les expériences que nous avons menées dans la section 4.4. Enfin, nous concluons et discutons de nos perspectives dans la section 4.5.

4.2 État de l'art sur l'apprentissage ensembliste

Dans cette section nous passons brièvement en revue quelques concepts de base de l'apprentissage ensembliste, ainsi que les travaux récents de la littérature ayant utilisé cette catégorie d'algorithmes pour la classification des rumeurs dans les média sociaux.

L'apprentissage ensembliste fait référence à la génération et à la combinaison de plusieurs modèles pour résoudre une tâche particulière d'apprentissage automatique. L'explication intuitive de la méthodologie ensembliste découle de la nature humaine. Souvent, la prise de décision par un groupe d'individus aboutit à un résultat plus précis, plus utile ou plus correct qu'une décision prise par un seul membre du groupe. C'est ce qu'on appelle généralement la sagesse de la foule ou du groupe [Surowiecki, 2005]. Ce concept peut être illustré par l'histoire de Sir Francis Galton (1822-1911), un philosophe et statisticien anglais qui a conçu le concept de base de l'écart-type et de la corrélation. Alors qu'il visitait une foire aux bestiaux, Galton a organisé un simple concours de devinette du poids d'un bœuf. Des centaines de personnes ont participé à ce concours, mais personne n'a réussi à deviner le poids : 1,198 *pound*. À sa grande surprise, Galton a découvert que la moyenne de toutes les suppositions était assez proche du poids exact. Dans cette expérience, Galton a révélé le pouvoir de combiner de nombreuses prédictions afin d'obtenir une prédiction plus proche de la réalité. Les méthodes ensemblistes illustrent ce concept dans les tâches d'apprentissage automatique où elles aboutissent à une amélioration des performances prédictives comparativement à un modèle unique. Il existe trois approches principales pour développer un modèle ensembliste [Zhang and Ma, 2012].

1. ***Boosting*** **:** La propriété clé des ensembles *boosting* est l'idée de corriger les erreurs de prédiction. Les modèles sont appliqués et ajoutés à l'ensemble de façon séquentielle, de sorte que le deuxième modèle tente de corriger les prédictions du premier modèle, le troisième corrige le deuxième modèle, et ainsi de suite. Cela implique généralement l'utilisation d'arbres de décision très simples, qui ne prennent qu'une ou quelques décisions, appelés *weak learners*. Les prédictions des *weak learners* sont combinées à l'aide d'un simple vote ou d'un calcul de moyenne, bien que les contributions soient pondérées proportionnellement à leur performance. L'objectif est de développer ce qu'on appelle un *strong learner* à partir de nombreux *weak learners* construits à cet effet. De nombreux algorithmes ensemblistes populaires sont basés sur cette approche, notamment *AdaBoost, Gradient boosting Machines* et *Stochastic Gradient boosting* (XGBoost et les

méthodes semblables).

2. ***Bagging*** (*Bootstrap AGGregatING*) **:** C'est une méthode d'apprentissage ensembliste qui cherche à obtenir un groupe diversifié de membres de l'ensemble en faisant varier les données d'apprentissage. Cela implique généralement l'utilisation d'un seul algorithme d'apprentissage automatique, presque toujours un arbre de décision non élagué et l'entraînement de chaque modèle sur un échantillon différent du même jeu de données d'entraînement. Les prédictions faites par les membres de l'ensemble sont ensuite combinées à l'aide de statistiques simples, telles que le vote ou le calcul de la moyenne. De nombreux algorithmes ensemblistes connus sont basés sur cette approche, notamment *Bagged Decision Trees, Random Forest* et *Extra Trees*.
3. ***Stacking*** **:** C'est une méthode d'ensemble qui recherche un groupe diversifié de membres en variant les types de modèles appliqués aux données d'apprentissage et en utilisant un autre modèle (*metalearner*) pour combiner les prédictions. Le *stacking* a sa propre nomenclature où les membres de l'ensemble sont appelés modèles de niveau 0 et le modèle qui est utilisé pour combiner les prédictions est appelé modèle de niveau 1. La hiérarchie à deux niveaux de modèles est l'approche la plus courante. Le *stacking* est probablement la technique de *metalearning* la plus populaire. En utilisant un *metalearner*, cette méthode tente d'induire quels classifieurs sont fiables et lesquels ne le sont pas. De nombreux algorithmes ensemblistes courants sont basés sur cette approche, notamment *Stacked Generalization, Blending Ensemble* et *Super Learner Ensemble*.

En faisant la moyenne ou le vote des résultats produits par l'ensemble des classifieurs, les méthodes ensemblistes fournissent de meilleures prédictions et évitent le surapprentissage (*overfitting*). Une autre raison qui contribue à la meilleure performance de l'apprentissage ensembliste est sa capacité à échapper aux minima locaux. En utilisant plusieurs modèles, l'espace de recherche s'élargit et les chances de trouver un meilleur résultat augmentent [Sagi and Rokach, 2018].

Aujourd'hui, les méthodes d'apprentissage ensemblistes ont montré de bonnes performances dans diverses applications, notamment la prédiction de l'irradiance solaire [Lee et al., 2020], l'analyse de la stabilité des pentes [Pham et al., 2021], le traitement du langage naturel [Sangamnerkar et al., 2020], la détection de logiciels malveillants [Gupta and Rani, 2020] et la détection des accidents de la circulation [Xiao, 2019]. Contrairement à d'autres applications, la classification des rumeurs à l'aide de techniques d'apprentissage ensemblistes a été très peu étudiée par les travaux de la littérature.

[Kaur et al., 2020] proposent un modèle de vote à plusieurs niveaux pour la tâche de détection des *fake news*. Ils utilisent trois techniques d'extraction de caractéristiques textuelles à savoir, TF-IDF, *Count-Vectorizer* (CV) et *Hashing-Vectorizer* (HV). L'étude conclut que le modèle de vote est plus performant que 11 autres modèles d'apprentissage automatique individuels.

Pour traiter le problème de la détection des fausses nouvelles, [Kaliyar et al., 2019; Abonizio et al., 2020] utilisent la technique ensembliste du *boosting* et comparent leur performance avec plusieurs modèles d'apprentissage automatique individuels. Les résultats sur plusieurs corpus montrent que le *boosting* obtient les meilleures performances que tous les autres modèles individuels.

Enfin, [Al-Ash et al., 2019] constatent que l'approche de *bagging* réalise des résultats meilleurs par rapport à trois algorithmes individuels d'apprentissage automatique, en l'occurrence, les SVM, *Multinomial Naive Bayes* (MNB) et les RF, pour la détection des rumeurs en utilisant un jeu de données en langue indonésienne.

Comme nous l'avons mentionné, le peu de travaux existants utilisent principalement les techniques de *boosting* ou de *bagging* basées sur la fusion des prédictions des modèles homogènes. Contrairement à ces travaux, nous proposons l'utilisation de l'approche de *stacking* par la fusion des prédictions de plusieurs modèles hétérogènes. Notre objectif est d'améliorer les performances de MONITOR en utilisant cinq algorithmes de *metalearning* dont les modèles de base sont les quatre modèles utilisés par MONITOR. Nous détaillons notre proposition dans la section suivante.

4.3 Modèles d'apprentissage ensembliste

Étant donné que les stratégies d'apprentissage ensemblistes, telles que le *bagging* et le *boosting*, font généralement appel à un seul algorithme d'apprentissage automatique (habituellement un arbre de décision), nous utilisons à la place la stratégie de stacking qui cherche un groupe diversifié de membres en variant les types de modèles. En effet, la diversité découle de la variété des modèles d'apprentissage automatique utilisés comme membres de l'ensemble. En tant que tel, il est souhaitable d'utiliser un ensemble de modèles qui sont appris ou construits de manières très différentes, ce qui garantit qu'ils font des hypothèses différentes et, par conséquent, ont des erreurs de prédictions moins corrélées. On peut résumer les éléments clés du stacking comme suit :

— un ensemble de données d'entraînement inchangé ;
— divers algorithmes d'apprentissage automatique (modèles de base) ;
— un modèle d'apprentissage automatique (métamodèle) pour apprendre comment combiner au mieux les prédictions.

La figure 4.1 illustre la stratégie de stacking de l'apprentissage ensembliste.

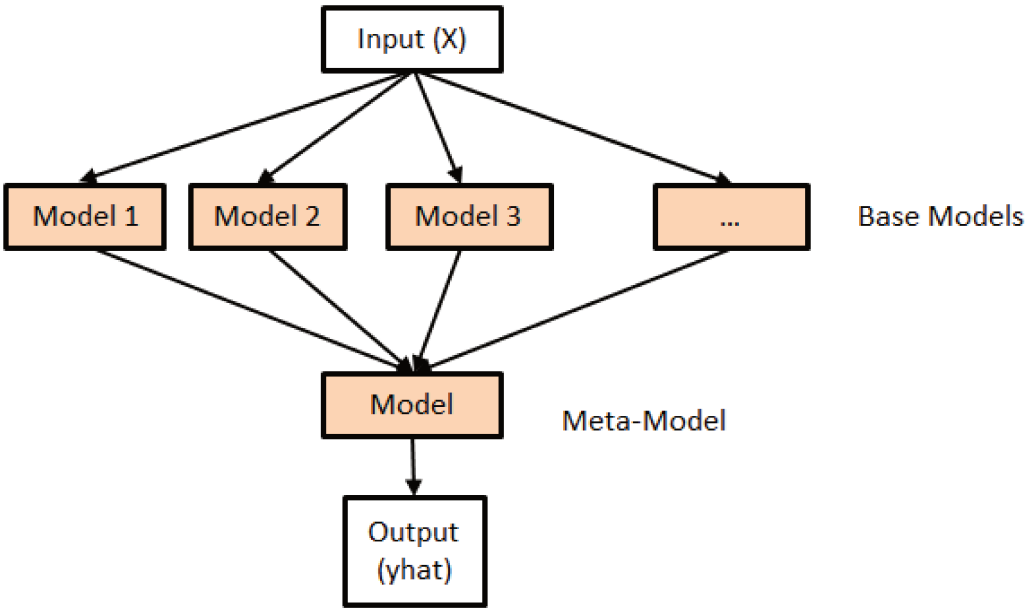


FIGURE 4.1 – Stratégie de stacking

Pour améliorer les performances de MONITOR, nous développons cinq algorithmes de *metalearning* comme variantes de la stratégie de stacking. Ces algorithmes utilisent les quatre classifieurs listés dans le dans le tableau 3.8 de la page 66 comme modèles de base. Il est à noter que chaque algorithme de base est évalué en utilisant ses hyperparamètres par défaut. Nous décrivons dans les sous-sections suivantes les cinq algorithmes de *metalearning* proposés.

4.3.1 Vote majoritaire

Le vote majoritaire (*voting ensemble*) est un métaclassifieur qui combine plusieurs prédicteurs de base en utilisant différentes règles de combinaison. Le prédicteur de base peut être n'importe quel modèle d'apprentissage automatique. Les prédicteurs de base individuels sont formés sur des données d'entraînement. La sortie de ces prédicteurs de base est combinée en utilisant certaines règles de combinaison comme la moyenne

des prédictions dans le cas de la régression et le vote majoritaire dans le cas de la classification.

Nous développons deux modèles de vote. Le premier est un modèle de *soft voting* appelé MONITOR$_{sv}$ qui additionne les prédictions faites par les modèles de classification de base et prédit le label de la classe avec la plus grande somme de probabilité. Le deuxième modèle est un modèle de *weighted average voting* appelé MONITOR$_{wav}$ où les votes des modèles de base sont proportionnels à leur performance. En effet, la performance de chaque modèle de l'ensemble sur les données d'entraînement est utilisée comme pondération relative du modèle lors des prédictions. La performance est calculée en utilisant la précision de classification comme un ratio de prédictions correctes compris entre 0 et 1. Un ratio avec des valeurs élevées signifie un bon modèle et, par conséquent, une plus grande contribution à la prédiction.

4.3.2 *Stacking* canonique

En reprenant la stratégie de stacking canonique (*Canonical Stacking Ensemble*) [Wolpert, 1992] (figure 4.1), nous construisons un modèle appelé MONITOR$_{st}$. Concrètement, nous utilisons trois répétitions d'une *stratified 10-fold cross valiation* sur les quatre modèles de classification pour préparer au métamodèle le jeu de données d'entraînement (prédictions). Nous choisissons la régression logistique comme métamodèle, car ce modèle linéaire simple est capable de fournir une interprétation homogène des prédictions faites par les modèles de base.

En outre, nous entraînons le métamodèle sur le jeu de données constitué à partir des prédictions obtenues suite à l'évaluation des quatre modèles de base, ainsi que sur le jeu de données d'entraînement original, en utilisant une *5-fold cross validation*. Cela vise à fournir un contexte supplémentaire au métamodèle pour mieux combiner les prédictions.

4.3.3 *Blending Ensemble*

Blending était le terme utilisé pour la stratégie de *stacking* lors du prix Netflix en 2009. Le prix impliquait des équipes qui cherchaient à prédire des recommandations de films plus performantes que l'algorithme natif de Netflix et un prix d'un million de dollars américains était attribué à l'équipe qui obtiendrait une amélioration de

prédiction de plus de 10 %. Le *blending* est une autre variante de la stratégie de *stacking* où, au lieu d'entraîner le métamodèle sur les prédictions issues des modèles de base sur les données d'entraînement, il est entraîné sur les prédictions générées par les modèles de base sur les données de test. C'est ainsi que les modèles de base sont entraînés sur les données d'entraînement et que le métamodèle est formé sur les prédictions faites par chaque modèle de base sur les données de validation.

Au moment où nous rédigeons ce manuscrit de thèse, la bibliothèque scikit-learn de Python [Pedregosa et al., 2011] ne prend pas en charge le *blending*. Nous avons donc implémenté un modèle de *blending* appelé MONITOR$_{bld}$ en utilisant des modèles de scikit-learn.

Pour implémenter notre modèle, nous commençons par diviser l'ensemble de données, d'abord en ensembles de formation et de test, puis l'ensemble de formation en un sous-ensemble utilisé pour former les modèles de base et un sous-ensemble pour former le métamodèle. Pour cela, nous utilisons une répartition 50/50 pour les ensembles de formation et de test, puis une répartition 67/33 pour les ensembles de formation et de validation. De plus, nous choisissons la régression logistique comme métamodèle (le *blender*) pour les mêmes raisons évoquées pour le *stacking* canonique.

4.3.4 *Super Learner Ensemble*

Un *super learner ensemble* est une configuration spécifique du *stacking* [Van der Laan et al., 2007]. Il s'agit d'abord de diviser les données d'entraînement en *k-folds*, puis d'évaluer tous les modèles de base en utilisant la *k-fold cross-validation*. Toutes les prédictions générées par chaque modèle sur les données de test de chacun des *k-folds* sont ensuite conservés et utilisés pour entraîner le métamodèle qui apprend comment combiner au mieux les prédictions. Nous résumons cette procédure dans l'algorithme 1.

En outre, la figure 4.2, qui est reproduite à partir de l'article original de Van der Laan et al. [2007], décrit le flux de données de ce modèle. Nous utilisons la bibliothèque Python MLENS [Flennerhag, 2017] pour implémenter le modèle ensembliste *super learner* appelé MONITOR$_{sl}$, où nous divisons les données d'apprentissage en $k = 10$ échantillons. Le nombre de modèles de base est toujours $m = 4$ (KNN, CART, SVM et RF).

Algorithm 1 Super learner ensemble

1: Select a k-fold split of the training dataset
2: Select m base models or model configurations
3: **for all** base models **do**
4: Evaluate using k-fold cross-validation
5: Store all out-of-fold predictions
6: Fit the model on the full training dataset and store
7: **end for**
8: Fit a metamodel on the out-of-fold predictions
9: Evaluate the model on a holdout dataset or use model to make predictions

4.4 Expérimentation

4.4.1 Protocole expérimental

Dans nos expériences, nous utilisons comme modèles de base les quatre modèles individuels d'apprentissage automatique KNN, CART, SVM et RF utilisés avec le *framework* MONITOR pour la classification de la véracité des messages. Ces modèles sont hétérogènes car leur fonctionnement est entièrement différent les uns des autres, et ils constituent de ce fait un bon groupe diversifié pour le *stacking*. Ces prédicteurs de base reçoivent en entrée les caractéristiques du texte, du contexte social et des images, qui ont été extraits des messages. Les prédictions produites par ces prédicteurs de base sont utilisées par les cinq modèles ensemblistes définis plus haut. Ces derniers classent la véracité des messages dans l'une des catégories suivantes : vrai ou faux.

Nous évaluons les modèles individuels de l'apprentissage automatique, ainsi que ceux de l'apprentissage ensembliste sur les deux jeux de données utilisés pour MONITOR, à savoir MediaEval et FakeNewsNet, dont les détails statistiques ont été déjà présentés dans le tableau 3.4 de la page 61. Pour rapporter les performances quantitatives de chaque modèle, nous utilisons les mêmes mesures de performance utilisées lors de l'évaluation de MONITOR, en l'occurrence, précision au sens d'*accuracy* (ACC), précision au sens de *precision* (Prec), rappel (Rec) et score F_1.

Le tableau 4.1 résume les résultats obtenus par les cinq algorithmes de *metalearning*. Nous comparons leurs performances avec le meilleur modèle individuel d'apprentissage automatique constaté par MONITOR, soit RF (tableau 3.8, page 66).

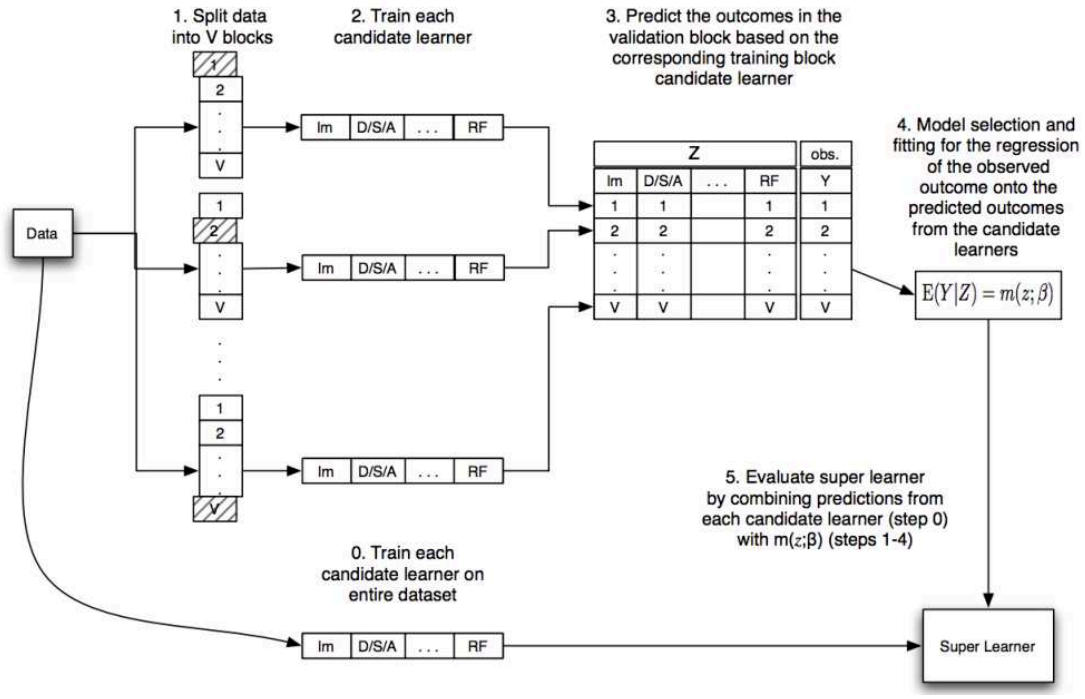


FIGURE 4.2 – Flux de données dans un algorithme *super learner* [Van der Laan et al., 2007]

4.4.2 Analyse des résultats

Notre analyse comparative des résultats expérimentaux montre que tous les modèles de méta-apprentissage sont plus efficaces que le meilleur modèle individuel d'apprentissage automatique (MONITOR), car en combinant plusieurs modèles, les erreurs d'un seul modèle de base sont probablement compensées par les autres modèles. Par conséquent, la performance globale de prédiction de l'ensemble est meilleure que celle d'un seul modèle de base.

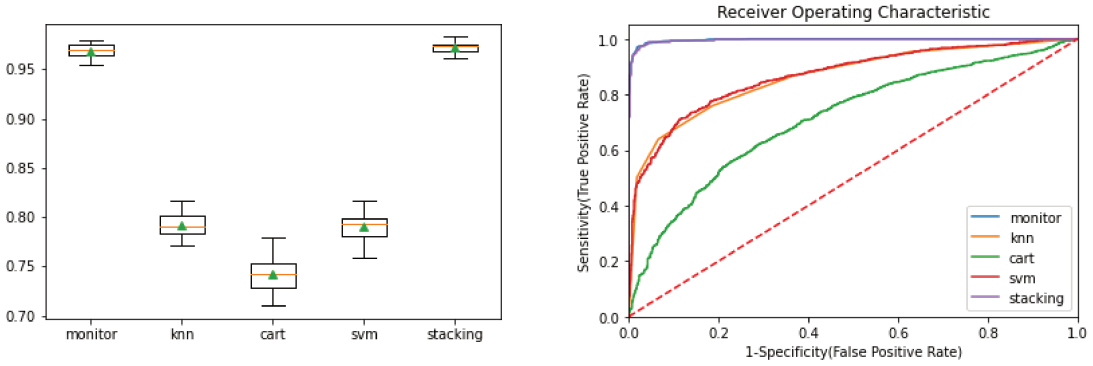
De plus, pour les deux jeux de données, l'algorithme de *stacking* canonique surpasse tous les autres modèles (selon toutes les mesures d'évaluation) avec 98.4 % et 93.6 % de précision (*accuracy*) sur MediaEval et FakeNewsNet, respectivement. Le modèle de *stacking* tire en effet avantage de la diversité des prédictions faites par les modèles contributeurs. Cela signifie que tous les algorithmes ont certaines compétences pour ce problème de classification, mais de manière différente.

Pour bien visualiser les performances des modèles de classification, nous utilisons la courbe ROC (*Receiver Operating Curve*). Le taux de vrais positifs (sur l'axe des y) est représenté par rapport au taux de faux positifs (sur l'axe des x). L'aire sous la

TABLE 4.1 – Performances de MONITOR et des modèles d'apprentissage ensembliste

Model	**MediaEval**				**FakeNewsNet**			
	Acc	**Prec**	**Rec**	$\boldsymbol{F_1}$	**Acc**	**Prec**	**Rec**	$\boldsymbol{F_1}$
MONITOR	0.962	0.965	0.966	0.965	0.889	0.914	0.864	0.889
MONITOR$_{sv}$	0.966	0.955	0.976	0.965	0.897	0.911	0.873	0.892
MONITOR$_{wav}$	0.968	0.968	0.970	0.969	0.906	0.90	0.927	0.914
MONITOR$_{st}$	**0.984**	**0.979**	**0.989**	**0.984**	**0.936**	**0.929**	**0.952**	**0.941**
MONITOR$_{bld}$	0.973	0.975	0.971	0.973	0.915	0.909	0.932	0.921
MONITOR$_{sl}$	0.970	0.980	0.959	0.969	0.921	0.915	0.937	0.926

courbe est considérée comme une métrique déterminante pour évaluer les performances de classification. Plus la mesure de l'aire sous la courbe est élevée, plus le modèle est capable de distinguer les faux messages. Les figures 4.3 et 4.4 illustrent la courbe ROC et le diagramme en boîtes du score de précision pour le modèle de *stacking* canonique comparativement aux algorithmes d'apprentissage automatique individuels (RF, CART, KNN et SVM) sur MediaEval et FakeNewsNet, respectivement.

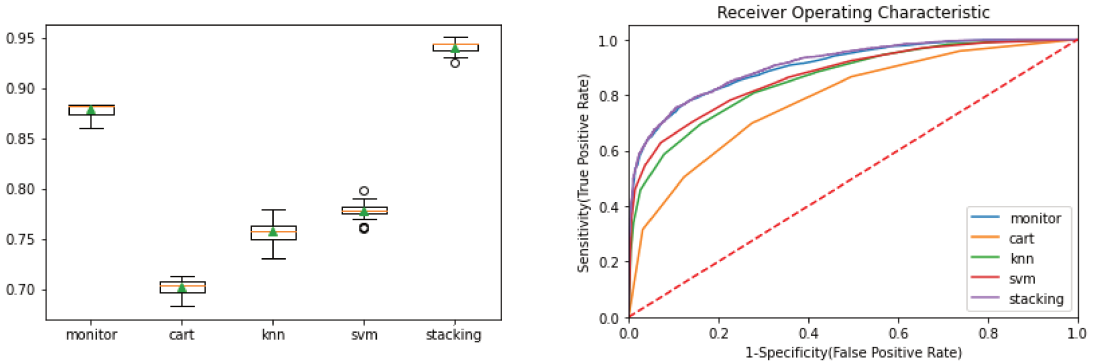


(a) Accuracy (b) ROC

FIGURE 4.3 – Modèle de *stacking vs.* modèles individuels sur MediaEval

Pour les deux jeux de données, nous pouvons observer que le *stacking* canonique est le meilleur modèle, puisque l'aire sous la courbe est la plus grande. D'autre part, l'aire sous les courbes ROC pour CART est la plus petite.

Parmi les cinq modèles ensemblistes, l'algorithme de *soft voting* MONITOR$_{sv}$ réalise les moins bons résultats, car il traite tous les modèles de base de la même façon, c'est-à-dire que tous les modèles contribuent de manière égale à la prédiction. Cette



(a) Accuracy (b) ROC

FIGURE 4.4 – Modèle de *stacking vs.* modèles individuels sur FakeNewsNet

hypothèse constitue une limitation pour ce modèle, car en réalité, les modèles de base ont des performances différentes.

Bien que l'algorithme de *stacking* canonique MONITOR$_{st}$ ait les meilleures performances, les algorithmes de *blending* MONITOR$_{bld}$ et de *super learner* MONITOR$_{sl}$ obtiennent des scores très proches de ceux du *stacking* et s'avèrent donc également très utiles pour la classification des rumeurs. Entre ces deux algorithmes, MONITOR$_{bld}$ est plus performant sur MediaEval, tandis que MONITOR$_{sl}$ est meilleur sur FakeNewsNet.

4.5 Conclusion

Les modèles traditionnels d'apprentissage automatique sont jusqu'ici largement utilisés pour la classification des rumeurs dans les réseaux sociaux. Cependant, les performances de ces derniers ne sont pas toujours probantes, car elles sont généralement liées aux caractéristiques du corpus utilisé. Dans ce chapitre, nous avons présenté une approche basée sur un autre paradigme d'apprentissage automatique, à savoir l'apprentissage ensembliste. Cette approche est complémentaire de MONITOR, car le but recherché est d'améliorer les performances de MONITOR en utilisant ce type d'algorithme pour la classification de la véracité des messages de *microblogs*.

À ce titre, cinq algorithmes d'apprentissage ensembliste ont été développés en se servant des quatre algorithmes (RF, CART, KNN et SVM) utilisés dans le *framework*

MONITOR comme modèles de base. Il s'agit des variantes de la stratégie de *stacking* (*soft* et *weighted average voting*, *canonical stacking ensemble*, *blending ensemble* et *super learner ensemble*).

Les expériences menées sur les jeux de données MediaEval et FakeNewsNet montrent que tous les modèles ensemblistes surpassent le meilleur modèle individuel d'apprentissage automatique (RF) sur les deux jeux de données. Plus important encore, l'analyse des métriques d'évaluation de la précision (*accuracy*) et le ROC, indiquent que, parmi les cinq modèles proposés, le *stacking* canonique est le meilleur modèle ensembliste sur les deux jeux de données. Cela n'exclut pas les performances réalisées par les algorithmes de *blending* et de *super learner* qui sont très proches de celles atteintes par le *stacking* canonique. Ces résultats confirment l'utilité des modèles ensemblistes pour la classification des rumeurs.

Pour le futur, nous avons deux directions de recherche. (1) Dans notre expérimentation, nous avons utilisé un nombre limité de modèles de base d'apprentissage traditionnel. Varier le nombre, le type et la combinaison des modèles de base, y compris des algorithmes d'apprentissage profond, pourrait améliorer encore les résultats de la classification. En effet, des travaux récents commencent à utiliser ce type de modèles ensemblistes. À titre d'exemple, Huang and Chen [2020] proposent un modèle ensembliste dont les modèles de base sont une combinaison de réseaux LSTM et CNN. Kaur et al. [2020] développent un modèle ensembliste avec plusieurs niveaux de vote en utilisant les prédictions de 12 modèles d'apprentissage automatique.

(2) MONITOR, ainsi que les modèles ensemblistes présentés dans ce chapitre, ont été évalués sur deux jeux de données du monde réel collectés sur Twitter. Si le jeu de données FakeNewsNet est relativement volumineux, ce n'est pas le cas de MediaEval (nombre de *tweets* et d'images limité). Ces propriétés ne permettent pas, par exemple, d'entraîner des modèles d'apprentissage profond connus pour leur besoin important en matière de ressources et de nombre d'instances de données.

Chapitre 5

deepMONITOR : un modèle multicanal basé sur l'apprentissage profond

Sommaire

Publication(s) associée(s)

Azri, A., Favre, C., Harbi, N., Darmont, J., Noûs, C. : Calling to CNN-LSTM for rumor detection : A deep multi-channel model for message veracity classification in microblogs. **ECML PKDD'2021.** European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (2021), Bilbao, Spain, September 13-17, 2021, Proceedings, Part 5, volume 12979, pages 497-513, Springer.

Dans les deux chapitres précédents, nous avons présenté le *framework* MONITOR et une approche basée sur les algorithmes d'apprentissage ensembliste pour améliorer ses performances. Ces techniques sont basées sur l'ingénierie des caractéristiques et les modèles d'apprentissage traditionnel. Dans ce chapitre, nous explorons les performances de l'apprentissage profond par la proposition d'un modèle neuronal profond pour la classification de la véracité des messages des *microblogs*, avec toujours l'idée d'une approche multimodale.

5.1 Introduction

Pour détecter les rumeurs sur les sites de *microblogs*, que nous ciblons particulièrement, la majorité des études existantes se concentrent sur les caractéristiques sociales disponibles dans les médias sociaux. Ces caractéristiques sont les métadonnées des messages, y compris les informations sur la façon dont les messages se propagent, par exemple, le nombre de *retweets*, de *followers*, de *hashtags* (#), les information de l'utilisateur, etc. Pour exploiter ces caractéristiques, de nombreuses solutions innovantes [Castillo et al., 2011; Ruchansky et al., 2017] ont été proposées. Malheureusement, ces caractéristiques ne sont pas toujours disponibles, par exemple, dans le cas où la rumeur vient d'être publiée et n'a pas encore été propagée. En outre, bien que les caractéristiques sociales soient utiles dans l'analyse des rumeurs, le contenu se révèle plus pertinent pour exprimer l'objectif de diffusion des rumeurs [Lin et al., 2015]. Par conséquent, dans ce chapitre, nous analysons le contenu des messages sous trois aspects afin de détecter automatiquement les rumeurs dans les *microblogs*.

Premièrement, les messages des médias sociaux ont un contenu textuel riche. Il est donc important de comprendre la sémantique d'un message pour détecter les rumeurs. Pour tenter d'automatiser la classification des messages comme vrais ou faux, on utilise généralement des techniques de traitement du langage naturel et d'apprentissage automatique qui s'appuient sur des caractéristiques textuelles spécifiques aux données, élaborées par l'ingénierie des caractéristiques [Castillo et al., 2011; Kwon et al., 2017]. Ces approches sont limitées, car les caractéristiques linguistiques des rumeurs varient selon les différents types de fausses nouvelles, d'événements et de plateformes de médias sociaux.

Deuxièmement, les images et les vidéos ont récemment gagné en popularité sur les *microblogs* et attirent une grande attention. Des données visuelles riches en informations

peuvent également être utiles pour classer les rumeurs. Toutefois, comme évoqué dans le chapitre 2, la prise en compte des images pour vérifier la véracité des messages n'est pas suffisamment explorée. Seules quelques études récentes exploitent le contenu multimédia [Jin et al., 2016b]. Cependant, ces caractéristiques sont conçues manuellement et peinent à représenter la distribution visuelle complexe de l'image.

Troisièmement, des travaux antérieurs ont montré que les fausses déclarations peuvent être détectées à partir du style d'écriture des auteurs et parfois être utilisées pour déduire leur personnalité [Pennebaker and King, 1999]. De plus, certains auteurs ont montré que les rédacteurs de fausses déclarations peuvent être détectés, car ils utilisent moins de marqueurs de complexité cognitive (ils racontent des histoires moins complexes), ils font moins de références à eux-mêmes pour se dissocier de l'information transmise et ont tendance à utiliser plus fréquemment des mots véhiculant des émotions négatives, comme un signe de culpabilité inconsciente [Newman et al., 2003]. Sur la base d'observations empiriques, Ajao et al. [2019] vérifient l'hypothèse selon laquelle il existe une relation étroite entre les rumeurs et les sentiments des textes postés en ligne. Dès lors, l'analyse des émotions à l'aide de méthodes de fouille d'opinion et d'analyse de sentiments peut aider à classer les rumeurs. L'analyse de sentiments à partir du texte généré par les réseaux sociaux sont des sujets de recherche bien étudiés. Cependant, à notre connaissance, l'analyse de sentiments dans le contexte de la détection de rumeurs dans les réseaux sociaux en ligne n'a pas été abordée jusqu'à présent par les travaux de recherche.

L'automatisation de la détection des rumeurs par rapport à l'une des trois caractéristiques mentionnées ci-dessus constitue déjà un véritable défi. Les caractéristiques textuelles et linguistiques créées par l'ingénierie des caractéristiques sont spécifiques aux données. Ces dernières prennent beaucoup de temps à produire. Les caractéristiques des images et les émotions, qui sont des indicateurs importants pour la détection des rumeurs dans les *microblogs*, ne sont toujours pas suffisamment étudiées.

Par ailleurs, les réseaux neuronaux profonds se sont révélés capables d'apprendre des représentations d'images et de textes bien plus précises que les descripteurs élaborés par l'ingénierie des caractéristiques. Ces modèles permettent d'améliorer les performances par rapport aux modèles d'apprentissage traditionnels en raison de leur capacité d'extraction automatique de caractéristiques. Plus particulièrement, les réseaux de neurones convolutifs (CNN) ont montré leurs puissantes capacités de représentation d'images, tandis que les réseaux de neurones récurrents (RNN) sont largement utilisés pour la représentation de texte. Inspirées par leurs succès, les techniques récentes d'apprentis-

sage multimodal utilisent les réseaux de neurones pour fusionner les caractéristiques multimodales dans de nombreuses applications, telles que la *visual question answering* [Antol et al., 2015] et l'*image captioning* [Vinyals et al., 2015].

Proposition et positionnement. Pour toutes ces raisons, nous proposons dans ce chapitre un modèle de bout en bout appelé deepMONITOR, qui est basé sur des réseaux de neurones profonds pour apprendre les représentations et fusionner les contenus textuels, les sentiments et les images des messages. L'originalité de notre approche réside particulièrement dans l'intégration, en plus du texte, dans l'apprentissage multimodal de la dimension visuelle et l'analyse des signes sentimentaux dans un contexte de détection de rumeurs. Ainsi, deepMONITOR peut exploiter des informations provenant de différentes modalités et capturer les dépendances sous-jacentes entre le contexte, les émotions et les informations visuelles d'une rumeur.

Concrètement, deepMONITOR est un modèle profond multicanal dans lequel nous employons d'abord un réseau de type *Long-term Recurrent Convolutional Network* (LRCN) pour capturer et représenter la sémantique des textes et les sentiments à travers des lexiques des sentiments. Cette architecture combine les avantages du CNN pour l'extraction des caractéristiques locales et la capacité de mémoire des réseaux *Long Short-Term Memory* (LSTM) pour bien connecter les caractéristiques extraites. Ensuite, nous utilisons le modèle pré-entraîné VGG-19 [Simonyan and Zisserman, 2014], fréquemment utilisé en vision par ordinateur pour extraire les caractéristiques visuelles saillantes des images attachées aux messages. Les caractéristiques des images sont ensuite fusionnées avec les représentations conjointes du texte et du sentiment pour classer les messages. Enfin, nous montrons expérimentalement que deepMONITOR surclasse les modèles de l'état de l'art de détection des rumeurs sur deux grands jeux de données multimédia collectés depuis Twitter, à savoir FakeNewsNet et DAT@Z21, qui est un jeux de données que nous avons construit à cet effet et que nous détaillons dans le chapitre suivant.

Ce chapitre est organisé de la manière suivante. Dans la section 5.2, nous décrivons formellement le modèle deepMONITOR. Ensuite, nous détaillons les résultats de l'évaluation quantitative et qualitative de notre modèle dans la section 5.3. Enfin, nous concluons ce chapitre et discutons des perspectives de recherche dans la section 5.4.

5.2 Modèle deepMONITOR

Pour répondre au problème délicat de la détection automatique des rumeurs, nous proposons d'exploiter toutes les modalités d'information des messages des *microblogs*. Dans cette section, nous définissons formellement le problème et introduisons quelques notations clés, puis nous présentons les composants de deepMONITOR.

5.2.1 Définition du problème et présentation du modèle

Nous définissons une instance de message comme étant $M = \{T, S, V\}$ composée d'une information textuelle T, d'une information de sentiment S et d'une information visuelle V. Nous désignons par C_T, C_S et C_V les représentations correspondantes. Notre objectif est d'apprendre une représentation discriminante C_M comme l'agrégation de T, S et V pour un message donné M, afin de prédire si M est un faux ($\hat{y} = 1$) ou un vrai message ($\hat{y} = 0$).

Dans la figure 5.1, nous présentons une vue d'ensemble de deepMONITOR. Tout d'abord, nous apprenons le texte avec un CNN, puis nous fusionnons la sortie avec un vecteur de sentiment avec deux LSTM empilés, ce qui génère une représentation conjointe C_{TS} pour ces deux modalités. La caractéristique visuelle C_V est obtenue avec le modèle pré-entraîné VGG-19. Enfin, C_{TS} et C_V sont concaténés pour former la représentation multimodale finale C_M du message M. C_M est l'entrée d'un classifieur binaire qui prédit si l'instance du message est fausse ou vraie. Dans la suite de ce chapitre, nous revenons sur chacuns des composants de ce modèle.

5.2.2 Réseaux LSTM

Par souci de clarté, nous commençons par présenter un bref rappel de la structure et du fonctionnement du modèle séquentiel LSTM. Le LSTM est un type spécial de réseau *feed-forward* RNN, qui peut être utilisé pour modéliser des informations séquentielles de longueur variable telles que des phrases ou des séries temporelles. La structure d'un réseau LSTM est illustrée dans la figure 5.2.

Étant donné une séquence d'entrée $\{x_1, x_2..., x_T\}$, un modèle RNN de base génère

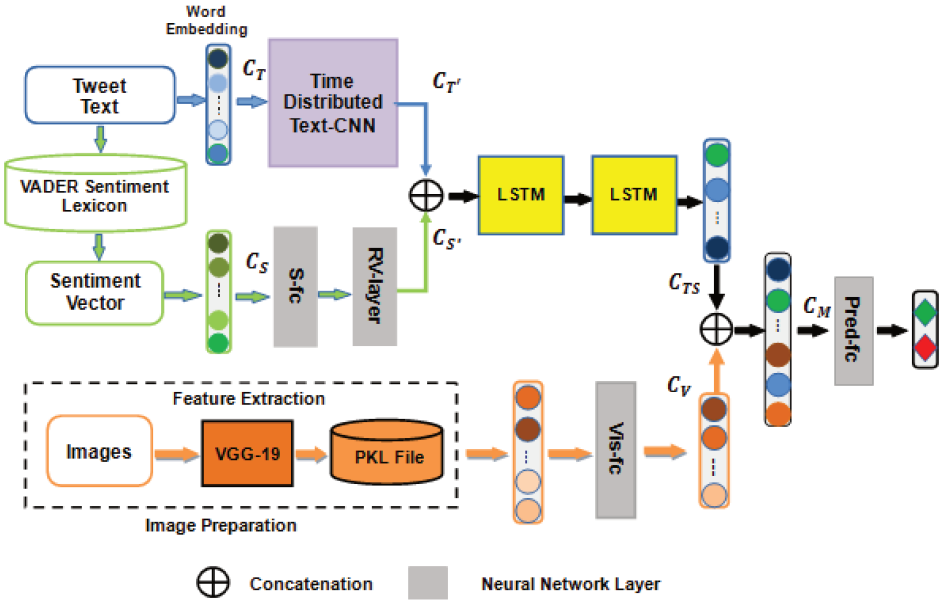


FIGURE 5.1 – Aperçu de deepMONITOR

la séquence de sortie $\{y_1, y_2..., y_T\}$, où T dépend de la longueur de l'entrée. Entre la couche d'entrée et la couche de sortie, il existe une couche cachée et l'état caché actuel h_t est estimé à l'aide d'une unité récurrente (équation 5.1).

$$h_t = f(h_{t-1}, x_t) \tag{5.1}$$

x_t est l'entrée actuelle, h_{t-1} est l'état caché précédent et f peut être une fonction d'activation ou une autre unité acceptant à la fois x_t et h_{t-1} comme entrée et produisant la sortie actuelle h_t.

Pour faire face à l'évanouissement ou à l'explosion des gradients [Bengio et al., 1994; Pascanu et al., 2013] dans l'apprentissage des dépendances temporelles à longue distance, les LSTM étendent les réseaux RNN basiques en stockant des informations sur de longues périodes de temps dans des unités de mémoire élaborées. Plus précisément, chaque cellule LSTM c est contrôlée par un groupe de vannes sigmoïdes : une vanne d'entrée i, une vanne de sortie o et une vanne d'oubli f qui se souvient de l'erreur pendant la propagation d'erreur [Hochreiter and Schmidhuber, 1997]. Pour chaque pas temporel t, la cellule LSTM reçoit des entrées de l'entrée courante x_t, de l'état caché précédent h_{t-1} et de la cellule mémoire précédente c_{t-1}. Ces portes sont mises à jour selon les équations 5.2–5.7 [Gers et al., 2002; Hochreiter and Schmidhuber, 1997].

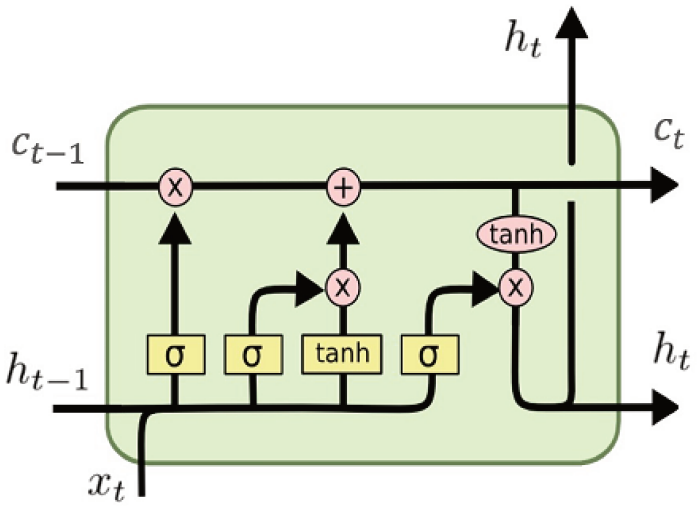


FIGURE 5.2 – Structure d'une cellule LSTM

$$i_t = \sigma(W_x^i x_t + W_h^i h_{t-1} + b_i) \tag{5.2}$$

$$f_t = \sigma(W_x^f x_t + W_h^f h_{t-1} + b_f) \tag{5.3}$$

$$o_t = \sigma(W_x^o x_t + W_h^o h_{t-1} + b_o) \tag{5.4}$$

$$\tilde{c_t} = tanh(W_x^c x_t + W_h^c h_{t-1} + b_c) \tag{5.5}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c_t} \tag{5.6}$$

$$h_t = o_t \odot tanh(c_t) \tag{5.7}$$

$W_\cdot^i$, $W_\cdot^f$ et $W_\cdot^o$ sont des matrices de poids pour les vannes correspondantes, et $b_\cdot$ sont des termes de biais qui sont appris du réseau. $\odot$ désigne la multiplication élémentaire entre deux vecteurs. σ est la fonction sigmoïde logistique. $tanh$ est la fonction tangente hyperbolique. La vanne d'entrée i décide du degré d'ajout d'une nouvelle mémoire à la cellule de mémoire. La vanne d'oubli f détermine le degré d'oubli de la mémoire existante. La cellule mémoire c est mise à jour en oubliant une partie de la mémoire existante et en ajoutant une nouvelle mémoire $\tilde{c}$.

5.2.3 Apprentissage des caractéristiques multimodales

5.2.3.1 Extraction de caractéristiques textuelles

Pour extraire des caractéristiques informatives de contenus textuels, nous utilisons un réseau CNN. Les CNN sont en effet efficaces dans de nombreux domaines. Nous incorporons un modèle CNN modifié, à savoir un Text-CNN [Kim, 2014], dans notre extraction de caractéristiques textuelles. L'architecture du Text-CNN est présentée dans la figure 5.3.

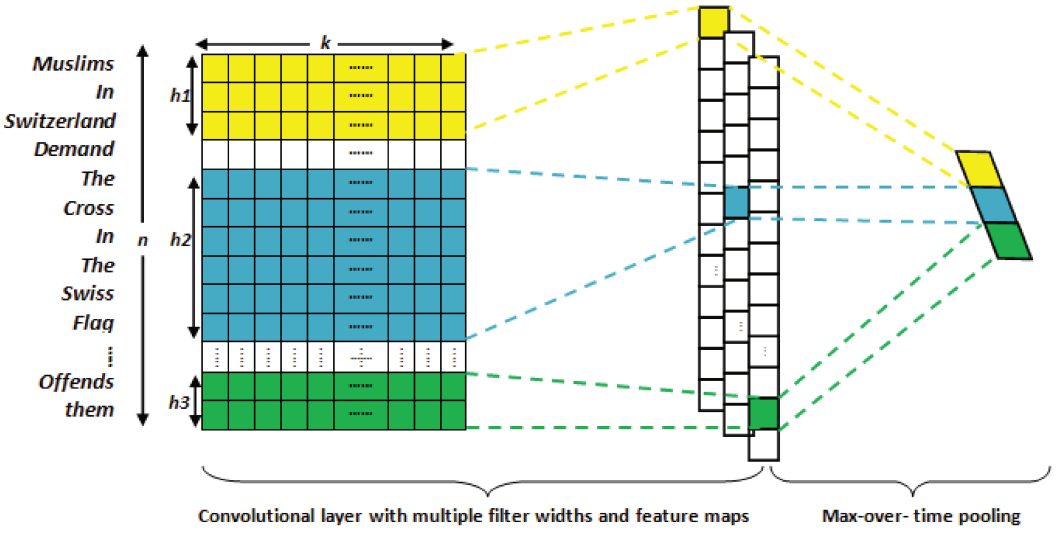


FIGURE 5.3 – Architecture du Text-CNN

Le Text-CNN bénéficie des avantages de plusieurs filtres avec différentes tailles de fenêtres, pour capturer différentes granularités de caractéristiques. Plus précisément, chaque mot du message est d'abord représenté comme un vecteur de prolongement de mots. Le vecteur de prolongement de chaque mot est initialisé avec un prolongement de mots pré-entraîné sur le jeu de données.

Étant donnée une instance de message contenant n mots, nous désignons par $T_i \in R^k$ le vecteur de prolongement de mots correspondant de dimension k pour le i^{e} mot du message. Le message est représenté par l'équation 5.8.

$$T_{1:n} = T_1 \oplus T_2 \oplus ... \oplus T_n \tag{5.8}$$

$\oplus$ est l'opérateur de concaténation. Pour produire une nouvelle caractéristique, un filtre de convolution avec une fenêtre de taille h prend en entrée la séquence contiguë de h mots dans le message. Par exemple, la caractéristique t_i générée à partir d'une fenêtre de taille h commençant par le i^e mot, est représentée par l'équation 5.9.

$$t_i = \sigma(W_c.T_{i:i+h-1} + b_c) \tag{5.9}$$

$W_c \in R^{h \times k}$ et $b_c \in R$ sont le poids et le biais du filtre, respectivement, et σ est la fonction d'activation linéaire rectifiée (ReLU). Ce filtre est appliqué à chaque fenêtre possible de h mots dans le message pour produire la carte de caractéristiques de l'équation 5.10.

$$t = [t_1, t_2, ..., t_{n-h+1}] \tag{5.10}$$

Ensuite, pour chaque vecteur de caractéristiques $t \in R^{n-h+1}$, nous appliquons une opération de *max-pooling* pour capturer les informations les plus importantes. Nous obtenons alors la caractéristique correspondante pour un filtre particulier. Le processus est répété jusqu'à ce que nous obtenions les caractéristiques de tous les filtres. Afin d'extraire les caractéristiques textuelles avec différentes granularités, différentes tailles de fenêtre sont appliquées. Pour une taille de fenêtre spécifique, nous avons d filtres différents.

Ainsi, en supposant qu'il y a c tailles de fenêtre possibles, nous avons donc $c \times d$ filtres au total. Après les opérations de *max-pooling*, une couche de *flatten layer* est nécessaire pour garantir que la représentation des caractéristiques textuelles $C_{T'} \in R^{c \times d}$ est réintroduite comme entrée dans le réseau LSTM.

Notons que le modèle Text-CNN ci-dessus permet de traiter plus d'un message, en le transformant de mots d'entrée en une représentation vectorielle interne. Nous voulons appliquer le modèle Text-CNN à chaque message d'entrée et transmettre la sortie de chaque message d'entrée au LSTM en une seule étape temporelle. Ainsi, nous devons répéter cette opération sur plusieurs messages et permettre à la couche suivante (LSTM) de construire un état interne et de mettre à jour les poids à travers une séquence de représentations vectorielles internes des messages d'entrée. Par conséquent, nous enveloppons chaque couche du Text-CNN dans une couche Time-Distributed [Keras,

2021b]. Cette couche permet d'obtenir le résultat souhaité en appliquant les mêmes couches plusieurs fois et en fournissant une séquence de caractéristiques de messages sur laquelle le LSTM peut travailler.

5.2.3.2 Extraction des caractéristiques de sentiment

Nous formons l'hypothèse que l'intégration de signaux émotionnels dans le modèle de classification des rumeurs doit apporter certains avantages. Pour extraire les signaux émotionnels des messages, nous adoptons une approche basée sur un lexique de sentiment, à savoir le *Valence Aware Dictionary and sEntiment Reasoner* (VADER), qui est un lexique et un outil d'analyse des sentiments basé sur des règles, spécifiquement adapté aux sentiments exprimés dans les médias sociaux [Hutto and Gilbert, 2014]. Ce modèle est sensible à la fois à la polarité (positive/négative) et à l'intensité (force) de l'émotion. VADER s'appuie sur un dictionnaire qui fait correspondre des caractéristiques lexicales à des intensités d'émotion connues sous le nom de score de sentiment. Le score de sentiment d'un texte peut être obtenu en additionnant l'intensité de chaque mot du texte. De plus, nous calculons certaines caractéristiques textuelles qui expriment une sémantique ou des sentiments spécifiques, comme les marques d'émotion (points d'interrogation et d'exclamation) et les émoticônes.

Nous formons la représentation initiale du sentiment $C_S = [s_1, s_2, ..., s_l]^T$, où l est la dimension des caractéristiques du sentiment et s_i est la valeur scalaire de la i^{e} dimension. Nous utilisons d'abord une couche entièrement connectée (couche S-fc dans la figure 5.1) pour produire une représentation appropriée du vecteur de sentiment $C_{S'}$ (équation 5.11).

$$C_{S'} = W_{sf} C_S \tag{5.11}$$

W_{sf} sont les poids dans la couche entièrement connectée. Ensuite, nous utilisons une couche Repeat Vector [Keras, 2021a] pour nous assurer que $C_{S'}$ a la même dimension (3D) que la représentation des caractéristiques textuelles $C_{T'}$.

Pour bien connecter les caractéristiques extraites, les représentations du sentiment et celles des caractéristiques textuelles sont ensuite concaténées et introduites dans deux réseaux LSTM empilés. L'empilement des couches cachées LSTM rend le modèle

plus profond, permet une représentation plus complexe de nos données de séquence et capture des informations à différentes échelles.

À chaque pas temporel i, le réseau LSTM prend en entrée $[C_{T_i'}, C_{S'}]$, c'est-à-dire la concaténation du i^e message $C_{T_i'}$ et de la caractéristique transformée de sentiment $C_{S'}$. La représentation conjointe résultante des caractéristiques du texte et du sentiment, désignée par $C_{TS} \in R^p$, a la même dimension (désignée par p) que la représentation visuelle des caractéristiques qui est abordée dans la sous-sous-section suivante. L'ensemble du processus est illustré dans la figure 5.4.

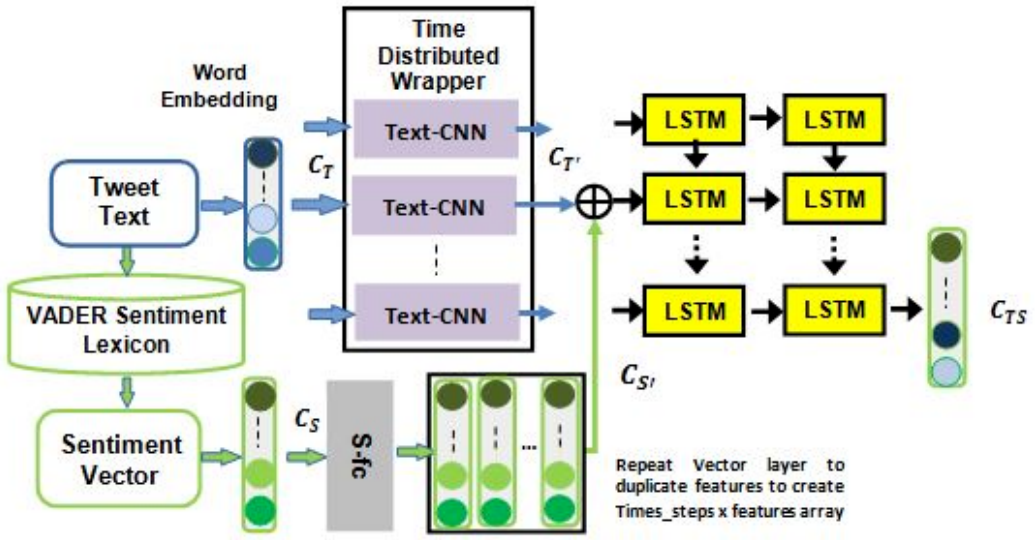


FIGURE 5.4 – Processus de fusion des caractéristiques du texte et du sentiment avec Text-CNN et deux LSTMs empilés

5.2.3.3 Extraction des caractéristiques de l'image :

Les images jointes aux messages constituent l'entrée du sous-réseau visuel (la branche inférieure de la figure 5.1). Nous utilisons le modèle pré-entraîné VGG-19 [Simonyan and Zisserman, 2014] pour générer des neurones visuels comme caractéristiques d'image. Nous conservons toutes les couches frontales du modèle VGG-19 et nous supprimons la dernière couche entièrement connectée, ainsi que la couche de classification.

Nous procédons à l'extraction des caractéristiques de toutes les images et nous les stockons dans des fichiers. L'avantage de cette méthode est qu'il n'est pas nécessaire de charger le très grand modèle pré-entraîné VGG-19, de le garder en mémoire et de

l'utiliser pour traiter chaque image pendant l'entraînement du sous-modèle textuel. Pour chaque caractéristique visuelle chargée, nous ajoutons une couche entièrement connectée (Vis-fc dans la figure 5.1) pour ajuster la dimension de la représentation finale de la caractéristique visuelle $C_V \in R^p$ (équation 5.12).

$$C_V = \psi(W_{vf} C_{V_{vgg}}) \tag{5.12}$$

$C_{V_{vgg}}$ est la représentation des caractéristiques visuelles obtenue à partir du modèle pré-entraîné VGG-19, W_{vf} est la matrice de poids de la couche entièrement connectée et ψ désigne la fonction d'activation ReLU. La représentation conjointe résultante des caractéristiques textuelles et de sentiment C_{TS} et la représentation des caractéristiques visuelles C_V sont ensuite concaténées pour former la représentation multimodale finale d'un message donné (équation 5.13).

$$C_M = C_{TS} \oplus C_V \in R^{2p} \tag{5.13}$$

5.2.4 Apprentissage du modèle

Jusqu'à présent, nous avons obtenu la représentation multimodale conjointe C_M d'un message donné M, qui est introduite dans une première couche entièrement connectée avec une fonction d'activation ReLu et une deuxième couche entièrement connectée avec une fonction d'activation sigmoïde pour prédire si les messages sont faux. La sortie de la couche sigmoïde pour le i^{e} message, désignée par $p(C_{M^i})$, est la probabilité que ce message soit faux (équation 5.14).

$$p(C_{M^i}) = \sigma(W_{df2} \psi(W_{df1} C_{M^i})) \tag{5.14}$$

W_{df1} et W_{df2} sont les poids dans les deux couches entièrement connectées, C_{M^i} est la représentation multimodale de l'instance du i^{e} message et σ et ψ sont les fonctions sigmoïde et ReLu, respectivement.

Nous utilisons la *cross-entropy* pour définir la fonction de perte de détection du i^{e} message (équation 5.15).

$$L(M^i) = -y^i \log p(C_{M^i}) - (1 - y^i) \log (1 - p(C_{M^i})) \tag{5.15}$$

y^i représente le label de vérité du terrain de l'instance du i^{e} message, 1 représentant les faux messages et 0 les vrais messages. Pour minimiser la fonction de perte, l'ensemble du modèle est entraîné de bout en bout à l'aide de la méthode *Batched Stochastic Gradient Descent* (équation 5.16).

$$L = -\frac{1}{N} \sum_{i=1}^{N} [y^i \log p(C_{M^i}) + (1 - y^i) \log (1 - p(C_{M^i}))] \tag{5.16}$$

N est le nombre total d'instances de messages. Notons que, pendant le processus d'optimisation, les paramètres des composants du modèle sont appris simultanément.

5.3 Validation expérimentale

Dans cette section, nous détaillons d'abord deux jeux de données de médias sociaux du monde réel utilisés dans nos expériences. Nous présentons ensuite les approches de détection des rumeurs de l'état de l'art, puis les détails de notre protocole expérimental. Enfin, nous analysons les performances de deepMONITOR par rapport aux méthodes existantes de la littérature.

5.3.1 Jeux de données

Pour fournir une évaluation des performances de deepMONITOR, nous menons des expériences sur deux jeux de données de médias sociaux du monde réel collectés sur Twitter. Les statistiques détaillées des deux jeux de données sont présentées dans le tableau 5.1.

FakeNewsNet [Shu et al., 2020] est l'un des référentiels les plus complets en matière de détection de fausses nouvelles. Les articles de fausses et de vraies nouvelles sont

TABLE 5.1 – Statistiques des deux jeux de données

Dataset	**FakeNewsNet**			**DAT@Z21**		
	Vrai	**Faux**	**Total**	**Vrai**	**Faux**	**Total**
Articles d'actualité	17 441	5 755	23 196	5 671	7 213	12 884
Avec images	17 214	1 986	19 200	1 765	2 391	4 156
tweets	1 042 446	565 314	1 607 760	1 209 144	1 655 386	2 864 530
Avec images	161 743	46 025	207 768	179 153	216 472	395 625
Images	163 192	49 582	212 774	211 447	271 419	482 866

collectés sur les sites web de vérification des faits PolitiFact [18] et GossipCop [19]. Les labels de vérité de terrain (faux ou vrai) des articles de nouvelles ont été fournis par des experts humains, ce qui garantit leur qualité. En plus du contenu des articles des nouvelles et des labels, FakeNewsNet fournit également des informations massives sur les utilisateurs du réseau social Twitter impliqués dans la diffusion de vrais et de faux messages. Puisque nous sommes particulièrement intéressés par les images dans cette thèse, nous extrayons et nous exploitons les informations relatives aux images de tous les *tweets*. Nous commençons par supprimer les images dupliquées et celles de mauvaise qualité. Nous supprimons également les *tweets* dupliqués et les *tweets* sans image, pour obtenir finalement 207 768 *tweets* avec 212 774 images jointes. Nous séparons soigneusement les données d'entraînement et de test de telle sorte que les *tweets* appartenant au même événement ne soient pas contenus à la fois dans les données d'entraînement et de test.

DAT@Z21 est un nouveau jeu de données que nous avons collecté sur Twitter. Plus concrètement, nous utilisons l'API du site web de vérification des faits Politifact pour récupérer les déclarations et les rapports de natures diverses vérifiés par des experts humains, en particulier les contenus publiés le 1er juin 2020. Nous utilisons les URL sources des pages web qui publient les articles d'actualités fournis par les professionnels de PolitiFact pour récupérer leur contenu. Pour garantir une vérité de base de haute qualité, nous ne retenons que les données et les métadonnées de 12 884 articles

18. https://www.politifact.com/
19. https://www.gossipcop.com/

explicitement étiquetés comme faux ou vrais. Pour extraire les *tweets* des utilisateurs qui discutent ou interagissent avec les articles, nous créons des requêtes avec les mots clés les plus représentatifs à partir du corps du texte et des titres des articles. Nous utilisons l'API de recherche académique de Twitter pour obtenir toutes les données correspondantes, notamment le contenu textuel, visuel, le contexte social (*retweets*, *reposts*, replies, etc.) et les profils des utilisateurs, en envoyant en arguments les requêtes préparées précédemment. Sur les 2 864 530 *tweets* collectés, nous supprimons les *tweets* contenant uniquement du texte et les images dupliquées pour obtenir 395 625 *tweets* avec 482 866 images jointes. Enfin, nous divisons l'ensemble des données en ensembles de formation et de test et nous nous assurons qu'ils ne contiennent aucun événement commun. La procédure de construction et les statistiques détaillées de ce jeux de données sont présentées dans le chapitre 6.

5.3.2 Protocole expérimental

Nous utilisons le modèle pré-entrainé de prolongement de mot GloVe (*Global Vectors for Word Representation*) [Pennington et al., 2014] après un prétraitement standard du texte. Notre choix s'est porté sur GloVe parce que le prolongement de mots est entraîné sur des *tweets*. Nous obtenons un vecteur de prolongement de mots à une dimension de $k = 50$ pour chaque mot dans les deux jeux de données.

Nous fixons le nombre de filtres du réseau Text-CNN à $d = 32$ et la taille de la fenêtre des filtres à $\{4, 6, 8\}$. Nous extrayons 14 caractéristiques de sentiment des deux jeux de données (Tableau 5.2). La taille cachée de la couche entièrement connectée des caractéristiques de sentiment est de 32. La représentation conjointe du texte et du sentiment utilise un premier LSTM de taille cachée 64 et un second LSTM de taille cachée 32.

Les caractéristiques de l'image sont issues de la sortie de l'antépénultième couche du modèle pré-entraîné VGG-19, pour générer un vecteur de 4096 dimensions. Ce vecteur est introduit dans une couche entièrement connectée avec une taille cachée de 32. La représentation finale des caractéristiques multimodales est introduite dans une couche entièrement connectée avec une taille cachée de 10.

deepMONITOR utilise une *batch size* de 64 instances. Dans nos expériences, chaque jeu de données a été séparé en 70 % pour les données d'apprentissage et 30 % pour les

TABLE 5.2 – Détails des caractéristiques du sentiment

Caractéristiques
Score Négatif/Positif/Neutre/Composé de Vader
Nbre de mots positifs/négatifs
Fraction de mots positfs/négatifs
Nbre d'émoticônes tristes/heureuses
Nbre de points d'exclamation/d'interrogation
Nbre de caractères majuscules
Nbre mots/caractères

données de test. Le nombre d'itérations est fixé à 100 dans la phase d'entraînement avec une stratégie d'*early stopping*[20] pour éviter le surapprentissage du modèle sur les deux jeux de données. Le *learning rate* est défini à 10^{-2}.

5.3.3 Méthodes de référence

Pour valider l'efficacité de deepMONITOR, nous le comparons à trois groupes de méthodes de référence : les méthodes monomodales, les méthodes multimodales et une variante de deepMONITOR.

5.3.3.1 Méthodes monomodales

Nous proposons trois approches simples, dans lesquelles les informations relatives au texte, au sentiment et à l'image sont utilisées séparément pour la classification des rumeurs.

— **Text** : deepMONITOR utilisant uniquement les informations textuelles.
— **Image** : deepMONITOR utilisant uniquement les informations visuelles.
— **Sent** : deepMONITOR utilise uniquement les informations relatives aux sentiments.

20. Arrêt de la formation du modèle lorsqu'une métrique d'évaluation a cessé de s'améliorer.

5.3.3.2 Méthodes multimodales

Nous comparons deepMONITOR avec deux méthodes récentes de l'état de l'art pour la détection multimodale de rumeurs.

— **att-RNN** [Jin et al., 2017a] est un modèle de réseau neuronal profond applicable à la détection multimodale des fausses nouvelles. Il utilise un réseau LSTM et le modèle VGG-19 avec un mécanisme d'attention pour fusionner et capturer les relations entre les caractéristiques visuelles et les caractéristiques textuelles et de contexte social. Nous avons défini les hyperparamètres tels qu'ils sont décrits dans Jin et al. [2017a] et nous excluons les caractéristiques du contexte social car le nombre de ces caractéristiques change en fonction des données.
— **SAFE** [Zhou et al., 2020b] est une méthode basée sur un réseau neuronal profond qui utilise les informations multimodales des articles d'actualité pour la détection des fausses informations, où la représentation des articles est apprise conjointement par les informations textuelles et visuelles, ainsi que leur relation (similarité). SAFE facilite la reconnaissance de la fausseté d'une nouvelle dans son texte, ses images, ou bien la « non-pertinence » entre le texte et les images.

Enfin, nous incluons également une variante de deepMONITOR dénommée deepMONITOR-, où les informations relatives au sentiment sont supprimées.

5.3.4 Analyse des performances

Nous présentons tout d'abord les performances générales de deepMONITOR en les comparant avec les méthodes de référence. Ensuite, nous effectuons une analyse des composantes en comparant deepMONITOR avec ses variantes. Enfin, nous analysons la partie LRCN du moddèle. Nos mesures d'évaluation sont la précision au sens d'*accuracy* (Acc), la précision au sens de *precision* (Prec), le rappel (Rec) et le score F_1.

5.3.4.1 Analyse des performances générales

Le tableau 5.3 montre les résultats expérimentaux de deepMONITOR et des méthodes de référence sur les jeux de données FakeNewsNet et DAT@Z21.

TABLE 5.3 – Comparaison des performances

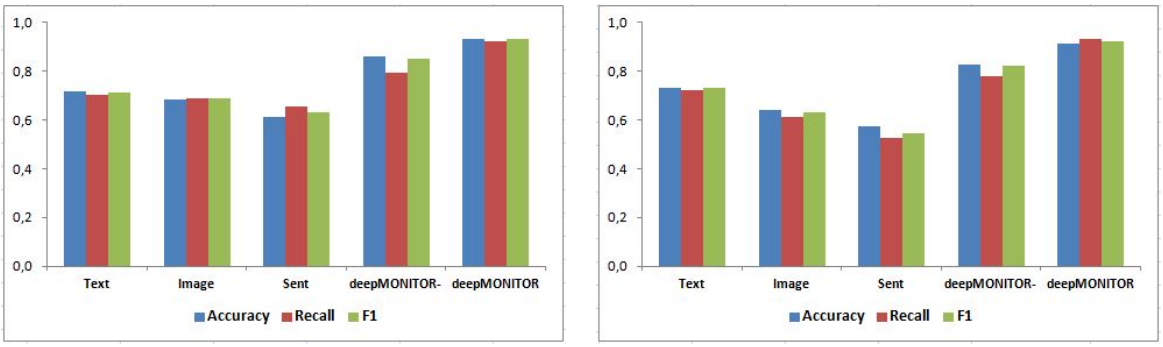
		Text	Image	Sent	deep MONITOR-	att-RNN	SAFE	deep MONITOR
Fake News Net	**Acc.**	0.715	0.686	0.612	0.862	0.786	0.883	**0.931**
	Prec.	0.725	0.685	0.60	0.820	0.774	0.861	**0.922**
	Rec.	0.702	0.688	0.655	0.796	0.810	0.938	**0.943**
	F_1	0.713	0.687	0.629	0.853	0.792	0.898	**0.932**
DAT @ Z21	**Acc.**	0.733	0.641	0.573	0.829	0.790	0.831	**0.913**
	Prec.	0.740	0.655	0.569	0.872	0.809	0.849	**0.905**
	Rec.	0.722	0.611	0.527	0.780	0.761	0.897	**0.932**
	F_1	0.731	0.632	0.547	0.824	0.763	0.861	**0.922**

Nous pouvons observer que les performances globales de deepMONITOR sont nettement supérieures à celles des méthodes de référence pour toutes les mesures de performance. De plus, les performances générales des méthodes multimodales forment un ordre : deepMONITOR > SAFE > att-RNN. deepMONITOR atteint en effet une précision globale de 93,1 % sur le jeu de données FakeNewsNet et de 91,3 % sur DAT@Z21, ce qui indique qu'il peut apprendre efficacement les caractéristiques conjointes de plusieurs modalités. Par rapport aux méthodes de référence, deepMONITOR améliore la précision (*accuracy*) de plus de 5 % et 8 % par rapport à SAFE et de 15 % et 12 % par rapport à att-RNN sur FakeNewsNet et DAT@Z21, respectivement.

5.3.4.2 Analyse des composantes

Les performances de deepMONITOR et de ses variantes sont présentées dans le tableau 5.3 et la figure 5.5. Les résultats laissent voir les observations suivantes.

1. L'intégration des informations textuelles, de sentiments et d'images des *tweets* donne les meilleurs résultats que ceux de toutes les variantes. Cela confirme que la combinaison de plusieurs modalités fonctionne plus efficacement pour la détection des rumeurs.
2. La combinaison des modalités textuelles et visuelles (deepMONITOR-) donne de meilleurs résultats que les variantes monomodales car, lors de l'apprentissage des informations textuelles, notre modèle utilise un réseau CNN avec plusieurs filtres et différentes tailles de fenêtres de mots. Comme la longueur de chaque



(a) FakeNewsNet (b) DAT@Z21

FIGURE 5.5 – Performances de deepMONITOR et ses variantes

message est relativement courte (moins de 240 caractères), le CNN peut capturer localement plus de caractéristiques représentatives, qui sont ensuite transmises aux réseaux LSTM pour connecter profondément les caractéristiques extraites.

3. Les performances obtenues avec les informations textuelles sont meilleures que celles atteinte avec les informations visuelles. Les caractéristiques textuelles sont en effet plus transférables et aident à capturer les motifs les plus partageables contenus dans les textes pour évaluer la véracité des messages. La raison est probablement que les deux jeux de données présentent une diversité de données suffisante. Ainsi, des caractéristiques linguistiques pertinentes peuvent être extraites pour la détection des rumeurs.
4. Les informations visuelles sont plus significatives que les informations de sentiments. Bien que la saisie de la sémantique des images constitue un véritable défi, l'utilisation du puissant outil VGG-19 permet d'extraire des représentations utiles des caractéristiques visuelles des images.
5. Les performances réalisées avec les informations de sentiments sont les plus mauvaises parmi les variantes multimodales, car sans contenu textuel ni visuel, le sens réel des *tweets* est perdu. Cependant, leur contribution n'est pas négligeable puisque l'utilisation de caractéristiques de sentiments (deepMONITOR- *vs.* deepMONITOR) permet d'améliorer la précision (*accuracy*) de 7 % et 8 % sur FakeNewsNet et DAT@Z21, respectivement.

5.3.4.3 Analyse des performances du réseau LRCN :

Dans cette section, nous analysons l'importance de la composante LRCN du modèle d'un point de vue quantitatif et qualitatif.

Analyse quantitative : À partir de deepMONITOR, nous développons deux nouveaux modèles, en supprimant le réseau Text-CNN dans le premier (deepMONITOR-CNN) et les deux réseaux LSTM dans le second (deepMONITOR-LSTM). Ensuite, nous exécutons les deux modèles sur le jeu de données FakeNewsNet. La figure 5.6 présente les résultats en terme de précision (*accuracy*), de rappel et de score F_1. La figure 5.6 montre que les performances de deepMONITOR sont meilleures que celles de deepMONITOR-CNN et deepMONITOR-LSTM pour les trois métriques.

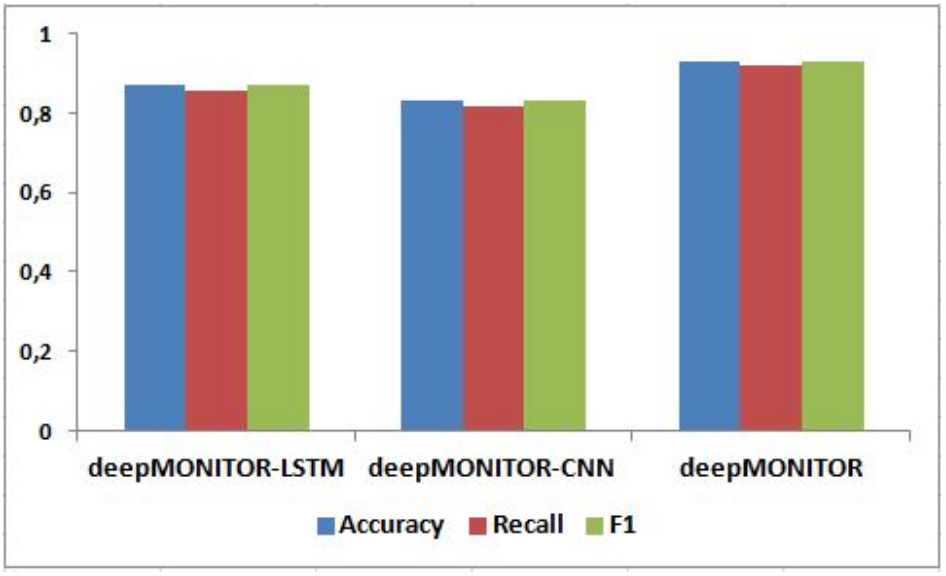
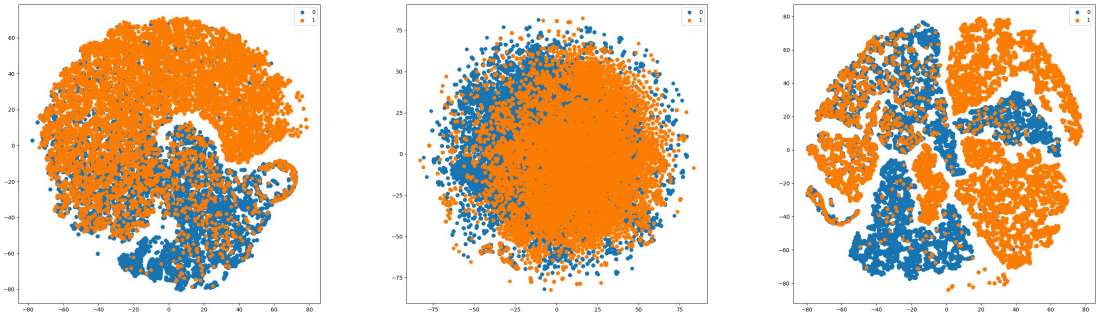


FIGURE 5.6 – Comparaison des performances des composants du réseau LRCN

Analyse qualitative : Pour analyser davantage l'importance de la composante LRCN de deepMONITOR, nous visualisons qualitativement la représentation des caractéristiques C_{TS} apprises par deepMONITOR, deepMONITOR-CNN et deepMONITOR-LSTM sur les données de test de FakeNewsNet avec t-SNE [Van der Maaten and Hinton, 2008] (figure 5.7).

Le label de chaque message est faux (couleur orange) ou vrai (couleur bleue). Nous pouvons observer que deepMONITOR-CNN et deepMONITOR-LSTM peuvent apprendre des caractéristiques discriminantes, mais les caractéristiques apprises sont

entrelacées. En revanche, les représentations des caractéristiques apprises par deepMONITOR sont plus discriminantes et nous voyons des zones de ségrégation plus grandes entre les deux labels. Cela s'explique par le fait que, lors de la phase d'entraînement du modèle, le Text-CNN peut extraire efficacement des caractéristiques locales et que les réseaux LSTM connectent et interprètent ces caractéristiques à travers les étapes temporelles. Nous pouvons donc en conclure que l'incorporation de la composante LRCN est essentielle et efficace pour la tâche de détection des rumeurs.



(a) deepMONITOR-LSTM (b) deepMONITOR-CNN (c) deepMONITOR

FIGURE 5.7 – Visualisations de représentations latentes du texte et des sentiments apprises sur les données de test de FakeNewsNet

5.4 Conclusion

Dans ce chapitre, nous avons présenté deepMONITOR, un nouveau modèle de bout en bout basé sur les techniques d'apprentissage profond pour la classification de la véracité des messages de *microblogs*. La majorité des approches de la littérature se concentrent sur les métadonnées des médias sociaux pour la classification des rumeurs. Malheureusement, ces caractéristiques sont spécifiques aux données et ne sont pas toujours disponibles. Contrairement à ces méthodes, deepMONITOR exploite le contenu des données des rumeurs, y compris les données visuelles des images et les informations de sentiments, qui n'étaient pas prises en considération jusqu'ici par les travaux de la littérature dans un contexte de détection de rumeurs. Considérant les caractéristiques multimodales dont disposent les *microblogs*, deepMONITOR permet d'extraire et de fusionner conjointement des informations textuelles, visuelles et de sentiments. Pour un message donné, nous fusionnons d'abord le texte et les signaux émotionnels avec un réseau LRCN. Ensuite, cette représentation conjointe est fusionnée avec les caractéristiques de l'image obtenues à partir du modèle pré-entraîné VGG-19.

Les expériences que nous avons menées sur deux grands jeux de données collectés sur Twitter ont démontré la pertinence de notre modèle. En effet, deepMONITOR réalise les meilleurs résultats par rapport aux techniques monomodales Text, Image et Sent, dans lesquelles les informations relatives au texte, aux images et au sentiments, respectivement, sont utilisées séparément pour la classification des messages. Notre modèle surpasse aussi deux techniques multimodales récentes de la littérature et la variante deepMONITOR-, où l'information des sentiments est supprimée. Nous avons pu ainsi valider empiriquement notre hypothèse de départ, à savoir que la prise en compte des modalités textuelles, visuelles et des signaux émotionnels des messages conduit à une classification plus précise de la rumeur.

Par ailleurs, nous avons montré l'utilité de la composante LRCN de notre modèle. Quantitativement parlant, deepMONITOR a obtenu les meilleurs performances que les deux variantes deepMONITOR-CNN et deepMONITOR-LSTM, en supprimant le Text-CNN et les deux LSTM, respectivement. Qualitativement parlant, la visualisation des représentations des caractéristiques apprises par deepMONITOR se sont révélées plus discriminantes que les deux autres variantes. Les résultats obtenus valident la pertinence de l'architecture du LRCN pour les problèmes qui ont une structure unidimensionnelle de mots dans une phrase, comme les messages de *microblogs*.

Une future ligne de recherche consiste à étudier plus en profondeur la contribution des caractéristiques de sentiment dans la détection des rumeurs. Dédier un sous-modèle profond pour l'apprentissage de telles caractéristiques, au lieu d'utiliser notre approche actuelle basée sur le lexique de sentiment, pourrait améliorer les performances de deepMONITOR.

Dans le chapitre suivant, nous détaillons la procédure de construction et de collecte du jeu de données DAT@Z21 utilisé pour la formation et l'évaluation du modèle deepMONITOR.

Chapitre 6

DAT@Z21 : Un jeu de données multimodal complet pour la classification des rumeurs

Sommaire

Publication(s) associée(s)

Le jeu de données DAT@Z21 est disponible publiquement à l'adresse suivante : https://git.msh-lse.fr/eric/dataz21.

Dans ce chapitre, nous présentons notre troisième contribution. Il s'agit de la construction d'un jeu de données qui comporte les propriétés essentielles pour la classification et l'analyse des rumeurs dans les *microblogs*. Ce jeu de données est partagé avec la communauté de recherche qui s'intéresse à la détection des rumeurs ou à l'apprentissage automatique d'une manière générale, et ce dans le strict respect des normes juridiques régissant les procédures de collection et de partage de données.

En plus des motivations ayant conduit à ce travail, évoquées dans la conclusion du chapitre 4, la construction de ce jeu de données constitue une étape préliminaire nécessaire pour valider les techniques d'apprentissage profond que nous avons abordées dans le chapitre précédent.

6.1 Introduction

Comme cela a déjà été discuté précédemment, de nombreuses solutions innovantes ont été proposées dans la littérature pour la détection des rumeurs. Nos contributions, présentées dans les chapitres 3, 4 et 5, font partie de cette démarche. Toutefois, ces techniques restent tributaires de la disponibilité de jeux de données de rumeurs complets et dotés de labels de qualité. Ceci constitue un problème pour la validation de la construction de nouvelles solutions pour une détection efficace des rumeurs, notamment pour les modèles d'apprentissage supervisé.

Détecter les rumeurs dans les médias sociaux est une tâche très difficile, car les fausses informations sont intentionnellement diffusées pour décevoir les utilisateurs, ce qui rend insuffisant leurs détection à partir de leurs seuls contenus. Ainsi, nous devons explorer des informations supplémentaires, telles que les engagements sociaux et les comportements sociaux des utilisateurs sur les plateformes de médias sociaux.

Les jeux de données disponibles sont peu nombreux et aucun d'eux ne fournit toutes les caractéristiques dont l'efficacité a été prouvée pour l'analyse des rumeurs, comme le contenu textuel, le contexte social, les informations visuelles et spatiotemporelles. La plupart de ces jeux de données ne fournissent que le contenu textuel ou les caractéristiques linguistiques [Wang, 2017; Nørregaard et al., 2019; Salem et al., 2019]. Cela représente un défi pour la construction des modèles de détection efficaces, car les caractéristiques linguistiques ne peuvent pas à elles seules refléter toutes les propriétés sous-jacentes des données de rumeurs.

Outre leurs propres limites, l'inconvénient commun des jeux de données existants est le manque de contexte social et d'autres informations utiles en dehors des données textuelles des articles d'actualité. En effet, très peu de jeux de données [Mitra and Gilbert, 2015; Tacchini et al., 2017; Shu et al., 2020] comportent également les informations du contexte social, qui représente la diffusion des articles d'actualité sur les plateformes des médias sociaux (messages et commentaires des utilisateurs). Cependant, ces jeux de données, qui fournissent l'engagement ou les interactions des utilisateurs sur les *microblogs*, ne couvrent généralement pas d'autres informations, telle que les profils de ces d'utilisateurs, à l'exception du jeu de données FakeNewsNet [Shu et al., 2020]. En conclusion, de par leur contenu, la majorité des jeux de données disponibles dans la littérature ne peuvent pas répondre aux défis posés par les rumeurs.

Proposition et positionnement : Soucieux du besoin de bénéficier d'un jeu de données qui réponde à ces exigences, nous proposons DAT@Z21, qui est, à notre connaissance, l'un des rares jeux de données volumineux et complet pour l'analyse des rumeurs dans les *microblogs*. DAT@Z21 est composé d'un ensemble d'articles de nouvelles incluant des labels de vérité terrain (faux ou vrai) de qualité fournis par les experts humains d'un site de vérification de faits. De plus, ces articles et labels de vérité terrain servent de base pour la collection des *tweets* qui représentent l'engagement social des utilisateurs qui diffusent ces articles sur Twitter.

Contrairement aux jeux de données existants, DAT@Z21 renferme toutes les informations nécessaires pour la détection des rumeurs, telles que le contenu textuel et linguistique, les informations spatiotemporelles, les engagements sociaux et les comportements des utilisateurs. De plus, notre jeu de données fournit également des contenus visuels qui jouent un rôle crucial dans le processus de diffusion des informations dans les *microblogs*.

Par ailleurs, nous apportons dans ce chapitre diverses analyses exploratoires (statistiques et distributions de données) sur des articles de presse, ainsi que sur leur diffusion sociale (*tweets*), afin de comprendre leurs principales propriétés et caractéristiques. Pour valider l'utilité de notre jeu de données, nous menons des expériences approfondies en utilisant plusieurs techniques de l'état de l'art pour la classification des rumeurs, en utilisant les données de DAT@Z21. Ces techniques sont obtenues en utilisant l'information monomodale ou multimodale des *tweets*, des algorithmes d'apprentissage automatique traditionnels et des modèles d'apprentissage profond.

Ce chapitre est organisé de la manière suivante. Dans la section 6.2, nous synthétisons les travaux de l'état de l'art relatifs aux jeux de données de détection de rumeurs. Dans la section 6.3 nous détaillons le processus de collection de données ainsi que les analyses et les statistiques exploratrices des données collectées. Nous présentons les expériences menées pour la classification des rumeurs en utilisant les données de DAT@Z21 dans la section 6.4. Enfin, nous concluons ce chapitre et discutons les perspectives de notre travail dans la section 6.5.

6.2 État de l'art sur les jeux de données

Dans cette section, nous passons brièvement en revue les jeux de données de détection et de classification des rumeurs, qui peuvent être regroupés essentiellement en deux catégories : (1) les ensembles de données récemment conçus pour les fausses informations de santé liées à la pandémie de la COVID-19 et (2) les jeux de données pour les fausses informations et les rumeurs en général.

6.2.1 Jeux de données sur la COVID-19

Récemment, avec l'émergence de la pandémie de la COVID-19, les rumeurs politiques et médicales autour de ce virus ont augmenté pour créer ce que l'on appelle communément l'infodémie globale. En peu de temps, de nombreux jeux de données sur la COVID-19 ont été publiés. La plupart de ces jeux de données n'ont pas d'annotation ou présentent des annotations automatisées produites par apprentissage par transfert ou par des méthodes semi-supervisées. Le contenu de ces jeux de données permet un suivi de l'avancement de cette pandémie en temps réel.

Les jeux de données collectés pour l'analyse de la COVID-19 fournissent généralement des informations sur les cas déclarés, l'estimation de la propagation du virus à partir de données épidémiologiques et démographiques. D'autres présentent des images médicales étiquetées de cas positifs et normaux de la COVID-19 ou du contenu textuel pour l'analyse des rumeurs relatives à la COVID-19 sur les médias sociaux.

Le premier jeu de données décrivant la pandémie de la COVID-19 à l'échelle mondiale

a été fourni par l'Université John Hopkins. Il s'agit d'un tableau de bord [21] en ligne et en temps réel, qui répertorie la localisation et le nombre de cas confirmés de la COVID-19, de décès et de guérisons pour tous les pays touchés [Dong et al., 2020]. Par ailleurs, Cui and Lee [2020] présentent CoAID, un jeu de données collecté pour l'analyse et la détection de rumeurs sur la COVID-19, qui contient des annotations automatiques pour les *tweets*, les réponses et les revendications relatives aux fausses nouvelles.

Zhou et al. [2020a] proposent ReCOVery, un jeu de données annotées de *tweets* partageant des nouvelles fiables ou non. Le jeu de données contient 2029 articles d'actualité sur le coronavirus publiés par 60 éditeurs de presse sources. La crédibilité de ces articles de nouvelles a été identifiée selon celle de leurs éditeurs, en utilisant des site spécialisés dans le référencement de la crédibilité des éditeurs de presse. Le jeu de données est également composé de 140 820 *tweets* qui indiquent la diffusion de ces articles sur Twitter. Enfin, FakeCovid [Shahi and Nandini, 2020] est un jeu de données multilingue qui contient 5 182 articles de nouvelles obtenus de 92 sites de vérification des faits. Les auteurs annotent manuellement les articles dans 11 catégories différentes en fonction de leur contenu.

6.2.2 Jeux de données de rumeurs

Un jeu de données de haute qualité joue un rôle extrêmement important dans la tâche de classification des rumeurs. Cependant, le manque de jeux de données labélisés sur les rumeurs constitue l'un des principaux obstacles pour le développement de techniques efficaces de détection des rumeurs. En général, les données d'actualité avec une vérité de terrain sont recueillies par des journalistes experts, des sites web de vérification des faits ou bien par le *crowdsourcing* [Shu et al., 2017]. Néanmoins, il n'existe pas de consensus sur un jeu de données de référence pour la détection des rumeurs. Cependant, certaines ressources disponibles publiquement méritent d'être mentionnées, dont la plupart ne présentent que des des articles de nouvelles, plus particulièrement des informations textuelles.

BuzzFeedNews [22] est un jeu de données qui contient les publications produites par neuf agences de presse vérifiées par Facebook au cours d'une semaine proche des élections américaines de 2016 (du 19 au 23 septembre et les 26 et 27 septembre). Chaque message

21. https://coronavirus.jhu.edu/map.html
22. https://www.buzzfeed.com/

et chaque article lié sont vérifiés manuellement et annotés pour s'assurer de leur véracité par cinq journalistes du site de vérification des faits BuzzFeed. Le jeu de données contient 1 627 articles de d'actualité annotés.

LIAR [23] [Wang, 2017] est collecté sur le site web de vérification des faits PolitiFact [24]. Il comprend 12 800 déclarations courtes recueillies sur le site PolitiFact let abélisées par des experts humains. Les déclarations sont classées en six catégories allant de complètement faux à complètement vrai.

BS Detector [25] est un jeu de données collecté à partir de l'extension de navigateur BS detector, développée pour vérifier la véracité des articles de nouvelles. Il recherche tous les liens d'une page web donnée pour détecter les références à des sources non fiables en les comparant à une liste de domaines compilée manuellement. Les labels sont les résultats de BS Detector, plutôt que ceux d'annotateurs humains.

NELA-GT-2018 [Nørregaard et al., 2019] est un jeu de données d'environ 714 000 articles d'actualité de sujets généraux, captés de février à novembre 2018. Les articles sont collectés à partir de 194 médias d'information avec de multiples labels obtenus directement auprès de plusieurs site de vérification de la crédibilité des sites web d'informations (NewsGuard, Pew Research Center, Wikipedia, OpenSources, MBFC, AllSides et BuzzFeed News). Ces labels font référence à la crédibilité des nouvelles, à la transparence, aux polarisations politiques et à l'authenticité.

FA-KES [Salem et al., 2019] est un jeu de données de fausses nouvelles sur la guerre en Syrie. Il se compose de 804 articles de nouvelles marqués comme vrais ou faux. Les labels ont été attribués à partir de la base de données du centre de documentation des violences syriennes, à l'aide d'un algorithme de *clustering*. Par conséquent, la fiabilité de ces labels peut poser problème.

L'inconvénient commun des jeux de données répertoriés ci-dessus est le manque de contexte social et d'informations autres que le texte des articles d'actualité. Outre le contenu des articles de nouvelles, d'autres chercheurs ont également collecté les caractéristiques de l'engagement des utilisateurs vis-à-vis de ces nouvelles, comme l'engagement des utilisateurs sur les médias sociaux en ligne.

23. https://www.cs.ucsb.edu/william/data/liardataset.zip
24. https://www.politifact.com/factchecks/
25. https://www.kaggle.com/mrisdal/fake-news

FakeNewsNet [26] [Shu et al., 2020] est un jeu de données collecté grâce aux sites de vérification des faits PolitiFact et GossipCop. Le contenu des articles d'actualité et les labels de vérité terrain correspondants sont extraits de ces deux sites. Les auteurs ont également collecté l'engagement social des utilisateurs par le biais de l'API de recherche avancée de Twitter. FakeNewsNet contient 1 056 articles de PolitiFact et 22 864 articles de GossipCop. Chaque article est étiqueté comme faux ou vrai. Au total, le jeu de données contient près de deux millions de *tweets*.

Enfin, CREDBANK [27] [Mitra and Gilbert, 2015] est un volumineux jeu de données qui contient environ 1000 événements d'actualité dont la crédibilité a été évaluée par 30 annotateurs issus de la plateforme web de *crowdsourcing Amazon Mechanical Turk* [28]. Le jeux de données compte environ 60 millions de *tweets* collectés entre 2015 et 2016.

Bien que ces jeux de données soient utiles pour la détection des rumeurs, ils ne couvrent généralement pas toutes les données utiles, telles que celles relatives aux profils des utilisateur, à l'exception de FakeNewsNet. En outre, les jeux de données existants se concentrent souvent uniquement sur le contenu textuel des articles d'actualité, et peu d'attention a été accordée au contenu visuel des articles d'actualité et à l'engagement des utilisateurs sur les médias sociaux. Or, les fausses images attachées aux messages peuvent facilement devenir virales sur les plateformes des médias sociaux et occasionner de graves perturbations sociales.

À la différence de ces jeux de données, nous proposons DAT@Z21, un jeu de données qui comprend non seulement des vrais et des faux articles d'actualité, mais aussi leur diffusion sur les médias sociaux. Nous recueillons toutes les caractéristiques d'intérêt possibles, y compris les informations multimodales (textuelles, visuelles et spatiotemporelles) et les données relatives à l'engagement des utilisateurs (profil de l'utilisateur, *tweets*, *retweets*, réponses, *likes*) (tableau 6.1). En outre, DAT@Z21 est mis à jour périodiquement pour collecter les derniers articles d'actualité et le contenu Twitter correspondant.

Après ce positionnement, nous explicitons dans la section suivante les étapes suivies pour la construction de notre jeu de données.

26. https://github.com/KaiDMML/FakeNewsNet
27. http://compsocial.github.io/CREDBANK-data/
28. https://www.mturk.com/

TABLE 6.1 – Comparaison de DAT@Z21 avec les jeux de données de rumeurs existants

Datasets	News Content		User Engagement					Spatiotemporal data	
	textual	Visual	User profil	Tweet	Retweet	Reply	Images	Spatial	Temporal
BuzzFeedNews	✓	—	—	—	—	—	—	—	—
LIAR	✓	—	—	—	—	—	—	—	—
BS Detector	✓	—	—		—	—	—	—	—
NELA-GT-2018	✓	—	—	—	—	—	—	—	—
FA-KES	✓	—	—	—	—	—	—	—	—
FakeNewsNet	✓	✓	✓	✓	✓	—	—	✓	✓
CREDBANK	✓	—	✓	✓	—	—	—	✓	✓
DAT@Z21	✓	✓	✓	✓	✓	✓	✓	✓	✓

6.3 Construction du jeu de données

Le processus global de collecte de données peut être divisé en trois principales étapes, comme le montre la figure 6.1 : (1) collecte des articles d'actualité avec des labels de vérité terrain, (2) préparation de la collecte de *tweets* et (3) collecte du contenu généré sur Twitter. Les sous-sections suivantes présentent ces différentes étapes.

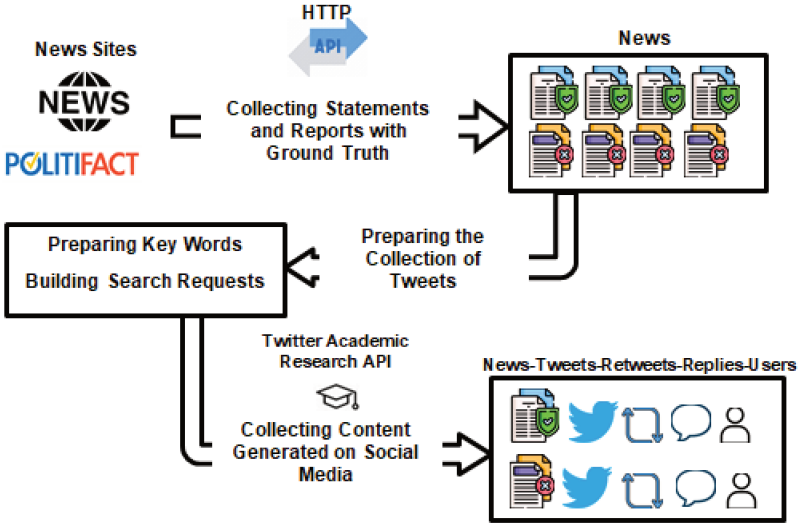


FIGURE 6.1 – Processus de collecte de données de DAT@Z21

6.3.1 Collecte des articles d'actualité et la vérité terrain

Les articles d'actualité en ligne peuvent être collectés à partir de différentes sources, telles que les pages web des agences de presse, les moteurs de recherche et les plateformes des médias sociaux. Cependant, la détermination manuelle de la véracité des informations

est une tâche difficile, qui nécessite généralement des annotateurs humains ayant une expertise du domaine, qui effectuent une analyse minutieuse des déclarations et des preuves supplémentaires, du contexte et des rapports provenant de sources fiables.

Une des sources disponibles pour obtenir une vérité de terrain de qualité sont les sites de vérification des faits. C'est ainsi que nous utilisons le site web de vérification des faits PolitiFact pour recueillir un ensemble d'articles de rumeurs vérifiés, annotés par de multiples labels de vérité terrain. Contrairement aux autres sites de vérification des faits, comme Snopes, où l'accès aux données est strictement interdit sans une autorisation préalable, PolitiFact dispose d'une API qui permet de collecter ses données dans les règles de l'art.

PolitiFact est un site web bien connu de vérification des faits de nature diverse. Pour chaque rumeur, des analystes professionnels fournissent la conclusion de la rumeur en termes de vérification, suivie d'une description complète de la source, de l'origine, des arguments qui soutiennent ou s'opposent à la rumeur, ainsi que de l'étiquette de véracité attribuée à la rumeur. La véracité de chaque rumeur peut être classée dans l'une des valeurs suivantes [29].

— **Vrai** : la déclaration est exacte et il ne manque rien de significatif.
— **Majoritairement vrai** : la déclaration est exacte, mais nécessite une clarification ou des informations supplémentaires.
— **À moitié vrai** : la déclaration est partiellement exacte, mais elle omet des détails importants ou présente des éléments hors contexte.
— **Majoritairement faux** : la déclaration contient un élément de vérité, mais ignore des faits essentiels qui donneraient une impression différente.
— **Faux** : la déclaration n'est pas exacte.
— **Ridicule** : la déclaration n'est pas exacte et constitue une affirmation ridicule.

Nous utilisons l'API Politifact [30] pour collecter les liens URL sources des pages web qui publient les articles d'actualité et les vérités de terrain correspondantes, fournies par les professionnels de PolitiFact. Après avoir identifié les pages web en question, nous y accédons à l'aide de la librairie Python *Beautiful Soup* [31].

Pour garantir une vérité de terrain de haute qualité et réduire le nombre de faux

29. Labels initiaux en anglais : *true, mostly true, mostly false, false, pants on fire !*
30. http://static.politifact.com/api/v2apidoc.html
31. https://pypi.org/project/beautifulsoup4/

positifs et de faux négatifs, nous sélectionnons uniquement les articles de nouvelles qui sont explicitement étiquetés comme vrais, faux ou complètement faux. Nous sommes particulièrement intéressés par les informations suivantes, liées à chaque article de nouvelles.

1. *Article ID :* l'identifiant unique attribués à chaque article.
2. *Article title :* le titre de l'article, qui peut bien résumer ou donner des indices sur le contenu de l'article.
3. *Article Author :* l'auteur ou les auteurs de l'article.
4. *Publication Date :* la date de publication de l'article.
5. *Article URL :* le lien URL de l'article.
6. *Article Bodytext :* le contenu textuel de l'article.
7. *Article Image :* le lien URL de l'image ou des images jointes à l'article.
8. *Article label :* la vérité terrain originale de la crédibilité de l'article, qu'elle soit vraie ou fausse.

Le tableau 6.2 résume les statistiques générales des articles d'actualité collectés dans cette première phase. Nous comptons 12 884 articles d'actualité dont un tiers (4 156 articles) comprend à la fois des informations textuelles et visuelles. La distribution des articles d'actualité par date de publication est présentée dans la figure 6.2. Nous pouvons voir que les dates de publication couvrent un large intervalle temporel (2007 à 2021). Cette information temporelle assure que notre collection d'articles d'actualité couvre divers sujets et événements de natures différentes tels que la politique, la santé, l'éducation, la science, la religion, l'immigration, etc.

TABLE 6.2 – Statistiques des articles d'actualité

	Vrai	Faux	Total
Articles d'actualité	5 671	7 213	12 884
Avec images	1 765	2 391	4 156

Les figures 6.3 et 6.4 montrent certaines caractéristiques textuelles du contenu des nouvelles générées à partir du titre et du corps du texte. À partir de la figure 6.3, nous pouvons noter que le nombre de mots dans les articles suit une distribution à longue traîne avec une valeur moyenne d'environ 513 et une médiane d'environ 418. En outre, la figure 6.4 montre le nuage de mots de tous les articles. Comme les articles d'actualité sont des déclarations, des rapports et des affirmations qui peuvent potentiellement être

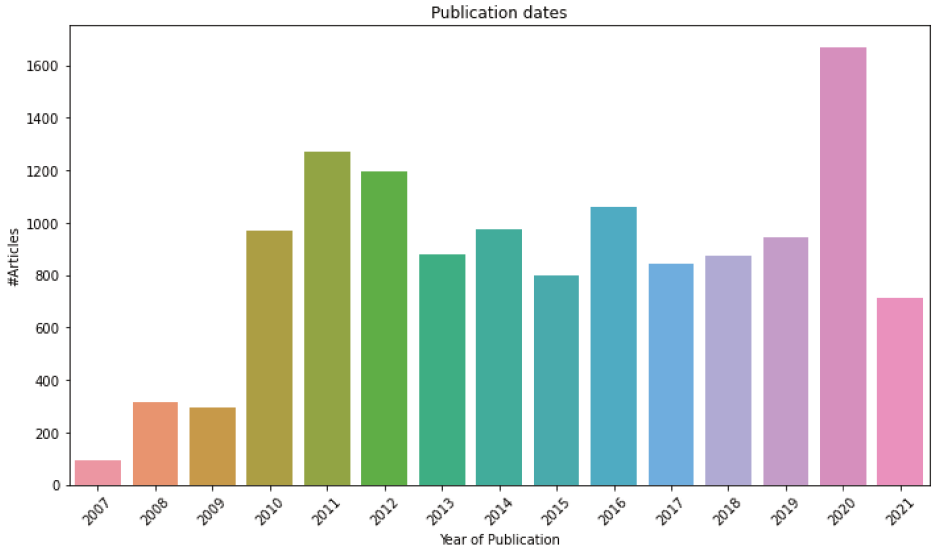


FIGURE 6.2 – Dates de publication des articles d'actualité

de fausses ou de vraies informations, il est naturel qu'ils partagent fréquemment des vocabulaires tels que *said* (11 131), *state* (10 112), *stated* (9 373) et *according* (9 208).

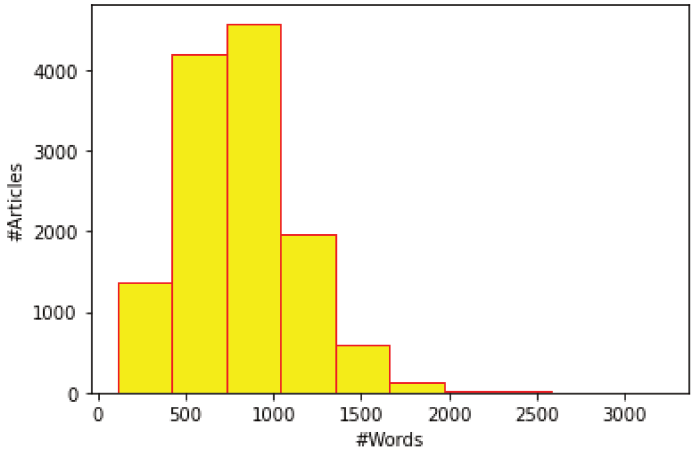


FIGURE 6.3 – Distribution du nombre de mots des articles d'actualité

En outre, à partir des titres des articles d'actualité, nous pouvons également saisir la source de la rumeur. Dans PolitiFact, cette source est généralement une déclaration d'un personnage de la classe politique, mais il peut également s'agir d'un message publié sur un média social ou un *blog*. La figure 6.5 illustre les dix principales sources d'articles relatives aux vraies et aux fausses nouvelles. Nous pouvons également remarquer que les publications sur Facebook, les images virales et les publications sur les *blogs* sont les trois principales sources d'articles relatifs aux fausses nouvelles.

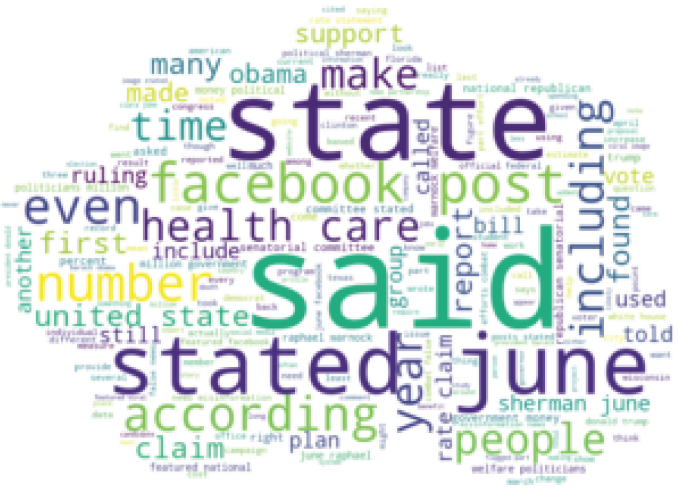
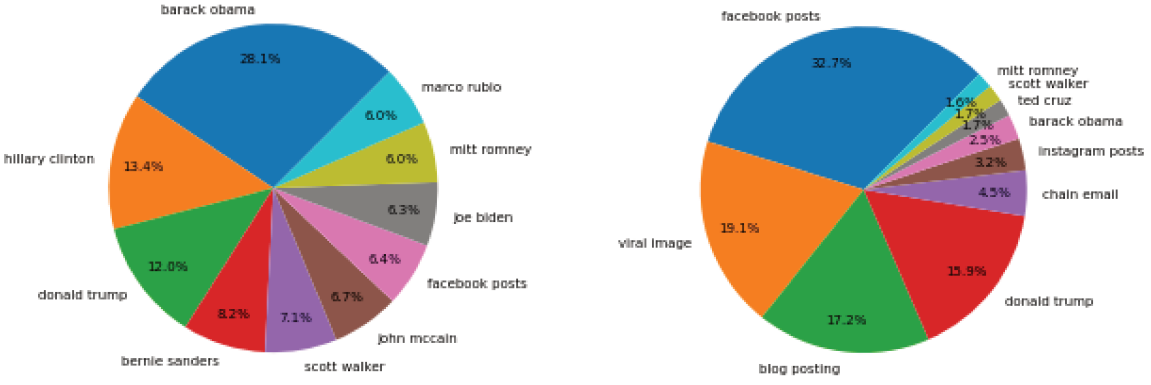


FIGURE 6.4 – Nuage de mots des articles d'actualité



(a) Vraies Nouvelles (b) Fausses nouvelles

FIGURE 6.5 – Dix principales sources de vrais et faux articles d'actualité

6.3.2 Préparation de la collecte de *tweets*

Dans cette seconde phase, nous suivons la diffusion des articles d'actualité sur Twitter en collectant les *tweets* des utilisateurs qui discutent ou interagissent avec ces événements. Il s'agit de chercher et d'extraire les mots-clés les plus représentatifs à partir du contenu de chaque article. Ces mots-clés servent de base à la construction des requêtes pour aller chercher le contenu Twitter associé aux articles. Notons que, dans cette étape, nous ne conservons que les articles d'actualité associés à des images. L'idée est de capturer des publications relayant probablement les images des articles ou d'autres images.

Nous construisons en premier lieu deux ensembles de mots-clés les plus représentatifs à partir du titre et du corps du texte de chaque article. Cette manière de procéder est motivée par la volonté de capturer tous les mots-clés figurant dans le titre et dans le corps du texte, car les titres utilisés par les journalistes ne représentent pas toujours le contenu des articles.

Nous développons un *pipeline* de traitement du langage naturel pour extraire les mots-clés du titre en supprimant les signes spéciaux et les mots vides courants (stop words). Dans cette étape, nous conservons au maximum six mots-clés pour éviter que les requêtes soient trop générales. Étant donné la taille relativement importante du corps du texte des articles, nous utilisons d'abord l'algorithme extractif de résumé de texte TextRank [Mihalcea and Tarau, 2004]. L'avantage d'utiliser un algorithme extractif plutôt qu'un algorithme abstractif est de conserver les mots-clés originaux du texte, ainsi que le sens ou le contexte général. Nous appliquons ensuite notre *pipeline* de traitement du langage naturel pour nettoyer le texte. Notons que, dans les deux ensembles de mots-clés (titre et corps du texte), nous conservons les entités nommées qui jouent un rôle important dans la sémantique des requêtes de recherche. Par exemple, la figure 6.6 montre les mots-clés extraits du titre et du corps du texte d'un article de presse.

6.3.3 Collecte des *tweets*

Jusqu'à présent, nous avons préparé deux ensembles de mots-clés qui servent de requêtes pour rechercher les *tweets* correspondants. Nous utilisons l'API de recherche

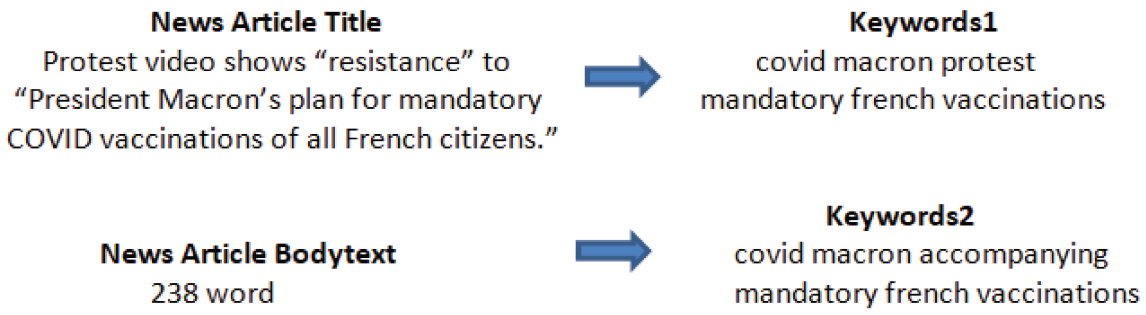


FIGURE 6.6 – Exemple d'extraction de mots-clés

académique de Twitter [32] avec le *full-archive search endpoint* pour collecter l'historique complet des *tweets* publics à la date du 7 septembre 2022. Il convient de préciser que cette API récente, qui présente des caractéristiques très avantageuses, est destinée aux chercheurs académiques ayant des objectifs de recherche spécifiques. L'accès à l'API de recherche académique est soumis à une présentation complète et détaillée du projet de recherche pour étude et décision des responsables de Twitter, une démarche que nous avons réalisée pour ce travail de thèse.

Avec une limite fixée à 1024 caractères pour les règles d'interrogation, ce *endpoint* prend également en charge un langage d'interrogation plus avancé (notamment, avec l'utilisation des opérateurs booléens) pour aider à produire des résultats plus précis, complets et non biaisés. Grâce à ces fonctionnalités, nous avons pu créer une seule requête de recherche qui combine les deux listes de mots-clés préparées précédemment, pour chaque article d'actualité.

En outre, le point final de recherche de l'archive complète renvoie les *tweets* correspondants avec des informations complètes telles que leurs identifiants, le texte, la date et l'heure de leur création. Il permet également d'extraire des détails précieux sur les utilisateurs, tels que leurs identifiants et leur activité sociale (nombre de *retweets*/*followers*/amis, contenu multimédia comme les images, géolocalisation et ensemble des métadonnées en utilisant les nouveaux paramètres d'expansion fournis par l'API de recherche académique).

Les détails statistiques de notre jeu de données ainsi constitué sont présentés dans le tableau 6.3.

32. https://developer.twitter.com/en/products/twitter-api/academic-research/product-details

TABLE 6.3 – Statistiques de DAT@Z21

Caractéristiques	**Vrai**	**Faux**	**Total**
Articles d'acualité	5 671	7 213	12 884
#Avec images	1 765	2 391	4 156
tweets	1 209 144	1 655 386	2 864 530
Avec images	179 153	216 472	395 625
Images	211 447	271 419	482 866
Re*tweets*	4 734	5 502	10 236
Replies	37 593	55 937	93 530
Likes	2 803	4 453	7 256
tweets avec location	3 823	4 497	8 320
Utilisateurs	98 967	115 669	214 636
Moyenne Followers	107 786	286 101	205 353
Moyenne Followees	2 929	3 153	3 051

Contrairement à la plupart des jeux de données existants dans la littérature, qui se concentrent uniquement sur le contenu textuel, nous voulons construire un jeu de données multimédia intégrant des images. Ainsi, nous collectons les textes originaux des *tweets* et les images jointes. Par conséquent, sur les 2 864 530 *tweets* collectés, nous supprimons les *tweets* contenant uniquement du texte, ainsi que les *tweets* dupliqués, pour obtenir un total de 395 625 *tweets* avec des images jointes. Cela représente une part importante (environ 14 %) du nombre total de *tweets*. En outre, parmi ces *tweets*, près de 10 % partagent plus d'une image. La figure 6.7 illustre la répartition des vrais et faux *tweets* en fonction du nombre d'images.

Pour le contenu textuel de notre jeu de données, nous analysons la distribution des thématiques des *tweets* portant une fausse ou vraie information. D'après la figure 6.8, nous pouvons observer que la distribution des thématiques des faux *tweets* est légèrement différente de celle des vrais. En effet, les deux distributions partagent la plupart des mots fréquents liés aux événements politiques comme *donald, trump, hillary, clinton* et *election* ou à la crise sanitaire de la COVID-19 comme *health, care, coronavirus* et *death*. La concordance des distributions des thématiques des vraies et fausses nouvelles garantit que les modèles de classification de rumeurs entraînés sur notre jeu de données ne sont pas des classifieurs de thématiques.

Pour les utilisateurs impliqués dans la diffusion de *tweets*, nous constatons que le nombre d'utilisateurs qui diffusent les fausses et les vraies nouvelles (214 636) est

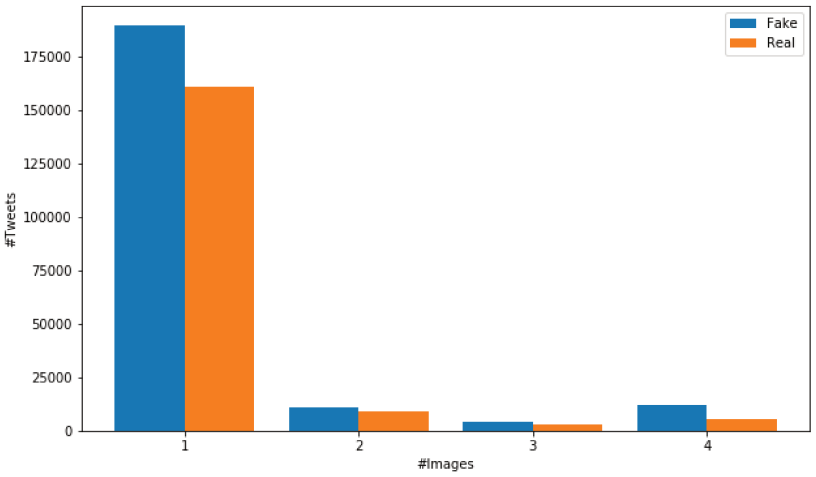
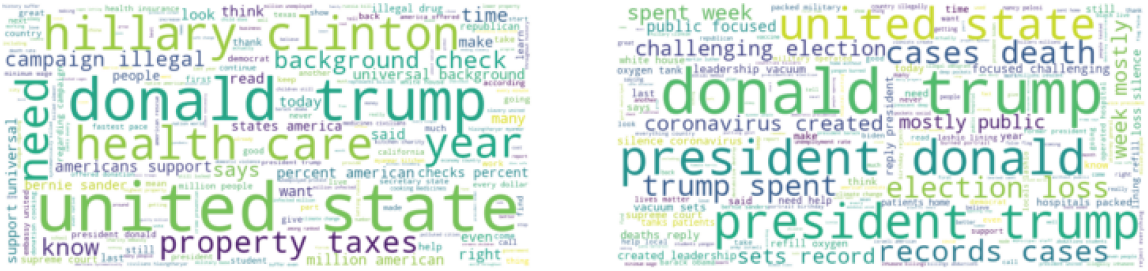


FIGURE 6.7 – Distribution des *tweets* en fonction du nombre d'images



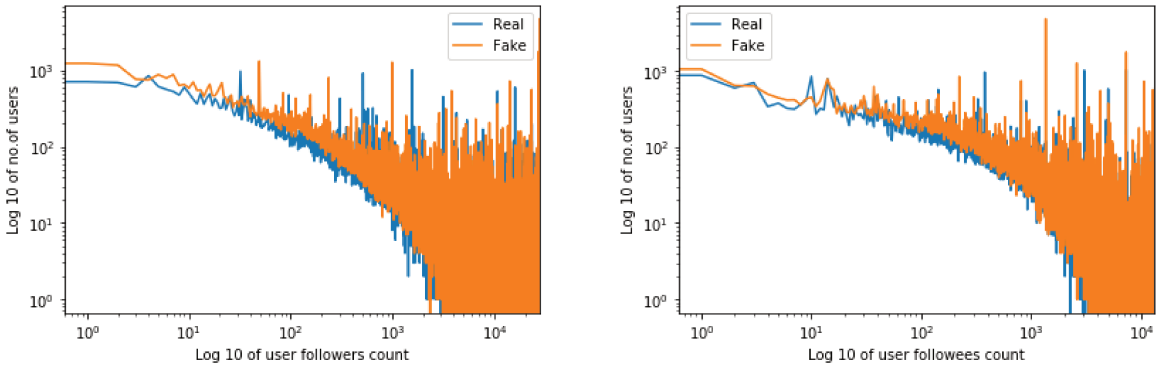
(a) Vrais *tweets*

(b) Faux *tweets*

FIGURE 6.8 – Distribution des thématique des faux et vrais *tweets*

supérieur au nombre total des utilisateurs distincts dans le jeu de données (194 692). Cela signifie qu'un utilisateur peut s'engager à la fois dans la diffusion de fausses et de vraies nouvelles. En ce qui concerne l'interaction sociale des utilisateurs, la distribution du nombre de *followers* et de *followees* des utilisateurs est présentée dans la figure 6.9. Nous pouvons observer que le nombre de *followers* et de *followees* des utilisateurs diffusant des faux et vrais *tweets* suit une distribution de type loi de puissance (*power-law*), ce qui est fréquemment observé dans les structures des réseaux sociaux.

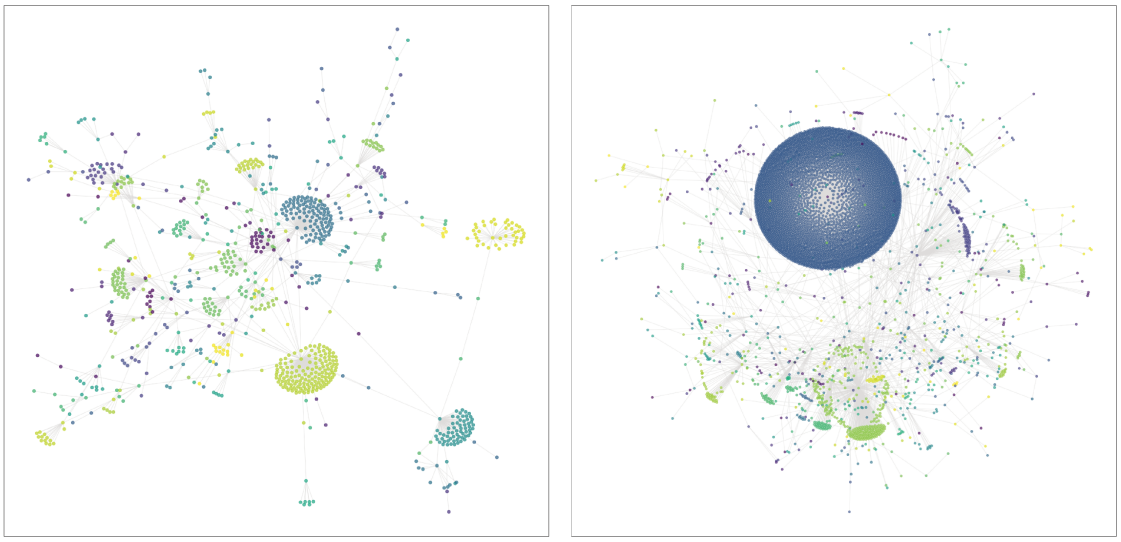
Pour comprendre comment les utilisateurs interagissent entre eux, nous créons un graphe d'interaction à l'aide de la librairie Python NetworkX [Hagberg et al., 2008]. Ce graphe représente trois types d'interactions entre deux utilisateurs : les *retweets*, les réponses et les *tweets* cités (*quoted tweets*). La figure 6.10 présente la plus grande composante connectée de notre graphe original pour les utilisateurs diffusant de vrais



(a) Nombre de *followers* (b) Nombre de *followees*

FIGURE 6.9 – Distribution du nombre de *followers* et de *followees*

et de faux *tweets*. Nous constatons que les deux utilisateurs présentent un cycle de chambre d'écho (*echo chamber* ; cf. section 1.3, page 12) et que les utilisateurs diffusant des faux *tweets* ont tendance à créer des groupes étroitement noués, caractérisés par des liens relativement denses.



(a) Vrais *tweets* (b) Faux *tweets*

FIGURE 6.10 – Graphe des connexions sociales pour les vrais et faux *tweets*

En outre, notre jeu de données comprend des informations spatiotemporelles qui décrivent l'engagement temporel des utilisateurs, ce qui fournit les informations nécessaires pour étudier plus profondément l'utilité des informations spatiotemporelles pour détecter les fausses nouvelles. La figure 6.11 illustre la distribution de la géolocalisation des utilisateurs qui publient des faux et des vrais *tweets*. Étant donné que la localisation

fournie explicitement par les utilisateurs dans leurs profils est généralement rare, nous prenons en compte les informations de localisation jointes aux *tweets*.

FIGURE 6.11 – Distribution spatiale des utilisateurs publiant de faux et de vrais *tweets*

Notons que, pour se conformer aux conditions d'utilisation de Twitter [33], il n'est pas autorisé de publier directement le contenu intégral des engagements sociaux et du réseau des utilisateurs. En revanche, nous pouvons publier les identifiants des *tweets* collectés pour permettre leur ré-hydratation (reconstitution du jeu de données) à des fins de recherche non commerciale, tout en fournissant les instructions pour obtenir les *tweets* à l'aide des identifiants publiés pour la convenance des utilisateurs. Pour plus de détails, le jeu de données est consultable à l'adresse `https://git.msh-lse.fr/eric/dataz21`. Nous comptons également assurer périodiquement la maintenance et la mise à jour de DAT@Z21 pour garantir sa pérennité.

6.4 Expérimentation

Dans cette section, nous menons des expériences comparatives sur la tâche de classification des rumeurs en utilisant le jeu de données DAT@AZ21. Les méthodes utilisées pour prédire la crédibilité des *tweets* peuvent être regroupées en fonction de leur schéma d'apprentissage, qui est soit un algorithme statistique traditionnel, soit un

33. `https://developer.twitter.com/en/developer-terms/agreement-and-policy`

réseau de neurones. Elles peuvent également être regroupées en méthodes monomodales ou multimodales. Tout d'abord, nous spécifions ces méthodes dans la sous-section 6.4.1. Ensuite, nous décrivons les paramètres expérimentaux dans la sous-section 6.4.2. Enfin, nous présentons une analyse des performances de ces méthodes dans la sous-section 6.4.3.

6.4.1 Méthodes de référence

6.4.1.1 Méthodes monomodales

***Naive Bayes* (NB) :** Nous utilisons le modèle de prolongement de document doc2vec pour représenter le texte et nous introduisons les représentations à un algorithme Gaussien de *Naive Bayes*.

***Logistic Regression* (LR) :** Nous utilisons le modèle de prolongement de mots word2vec pour représenter le texte et nous introduisons les représentations au modèle.

***Random Forest* (RF) :** Nous utilisons TF-IDF (*Term Frequency-Inverse Document Frequency*) pour représenter le texte et nous le soumettons au modèle.

Text-CNN [Kim, 2014] est un réseau de neurones convolutif pour la classification de texte. Nous introduisons le prolongement de mots du texte dans un réseau convolutif unidimensionnel (Conv1D) qui apprend des filtres. Ensuite, nous ajoutons une couche de *max pooling*.

6.4.1.2 Méthodes multimodales

SAFE[34] [Zhou et al., 2020b] est une méthode basée sur un réseau neuronal qui utilise des informations multimodales pour détecter les fausses nouvelles, où la représentation des nouvelles est apprise conjointement par leurs informations textuelles et visuelles,

34. https://github.com/Jindi0/SAFE

ainsi que leur relation de similarité. SAFE facilite la reconnaissance de la fausseté d'une nouvelle dans son texte, ses images ou la non-pertinence entre le texte et les images.

att-RNN [Jin et al., 2017a] est un modèle profond qui utilise un réseau LSTM et le modèle pré-entrainé VGG-19 avec un mécanisme d'attention pour fusionner les caractéristiques textuelles, visuelles et le contexte social des *tweets*. Nous définissons les hyperparamètres comme dans Jin et al. [2017a].

deepMONITOR [Azri et al., 2021a] est notre proposition, décrite dans le chapitre précédent.

6.4.2 Paramètres expérimentaux

Nous avons divisé aléatoirement le jeu de données DAT@Z21 en sous-ensembles d'entraînement et de test avec un ratio de 0,7/0,3. Nous utilisons la bibliothèque Python scikit-learn Python pour implémenter les algorithmes traditionnels avec les paramètres par défaut et sans réglage, et la bibliothèque Pythin Keras[35] pour implémenter les méthodes multimodales. Nous utilisons le modèle TF-IDF avec 1000 caractéristiques, les prolongements word2vec et doc2vec avec un vecteur de 200 dimensions pour représenter le texte.

Comme notre jeu de données a une distribution légèrement déséquilibrée entre les classes fausses et vraies (environ 1,17/1), nous utilisons les métriques suivantes pour évaluer la performance des algorithmes de classification des rumeurs : *Area Under Precision-Recall Curve* (PR AUC), *Area Under Receiver Operating Characteristic Curve* (ROC AUC), précision (*precision*), rappel et score F_1. Enfin, nous exécutons chaque méthode trois fois et nous reportons le score moyen dans le tableau 6.4.

6.4.3 Résultats expérimentaux

Le tableau 6.4 montre que, parmi les quatre modèles monomodaux, Text-CNN est le plus performant et obtient les meilleurs scores pour quatre des cinq mesures de

35. https://keras.io/

TABLE 6.4 – Performances de classification des rumeurs avec DAT@Z21

Type	Methods	PR AUC	ROC AUC	Prec.	Rec.	F_1
Monomodal	**LR**	0.778	0.799	0.729	0.777	0.752
	NB	0.496	0.561	0.593	0.480	0.531
	RF	0.843	0.846	0.754	0.783	0.764
	Text-CNN	0.861	0.866	0.759	0.774	0.771
Multimodal	**att-RNN**	0.880	0.871	0.805	0.753	0.777
	SAFE	0.937	0.935	0.851	0.857	0.854
	deepMONITOR	0.984	0.982	0.901	0.948	0.924

comparaison, à savoir le PR AUC, le ROC AUC, la précision et le score F_1. Pour les techniques multimodales, deepMONITOR surclasse les autres modèles avec les meilleurs scores dans toutes les métriques de comparaison.

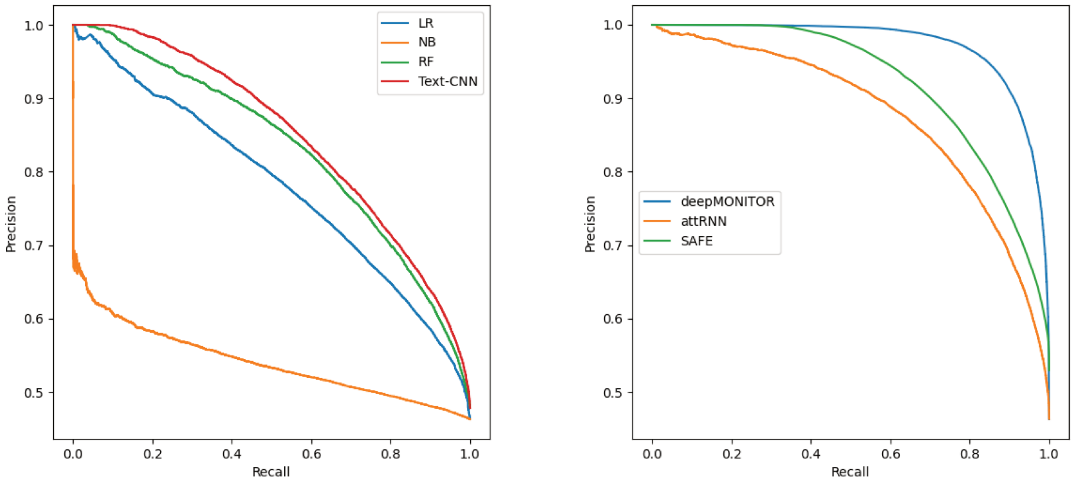
Les méthodes multimodales obtiennent généralement de meilleurs résultats comparativement aux méthodes monomodales, car elles incorporent des signaux à partir d'autres modalités de message que le contenu textuel, comme des informations visuelles, de sentiment ou de contexte social, ce qui permet de mieux capturer les informations sémantiques des *tweets*.

Comme l'ensemble de données est légèrement déséquilibré en faveur des faux labels, les modèles ont tendance à générer un peu plus de cas de faux positifs. Toutefois, les valeurs de la PR AUC, du rappel et du score F_1 sont satisfaisantes. La figure 6.12 illustre la courbe PR AUC pour les méthodes monomodales et multimodales.

Étant donné que la proportion de vraies et de fausses informations est susceptible d'être encore plus déséquilibrée dans les données du monde réel, les solutions de détection et de classification des rumeurs doivent gérer plus efficacement ce type de problème.

6.5 Conclusion

Dans ce chapitre, nous avons construit un jeu de données multimodales et complet appelé DAT@Z21, destiné à la classification des rumeurs dans le *microblog* Twitter. Ce jeu de données se distingue des jeux de données existants d'une part, par la qualité



(a) Méthodes monomodales

(b) Méthodes multimodales

FIGURE 6.12 – Courbe PR AUC pour les méthodes monomodales et multimodales

de vérité terrain collectée auprès des experts humains d'un site de vérification de faits. D'autre part, DAT@Z21 présente un ensemble de données variées et riches en information nécessaires pour la classification des fausses nouvelles, comme le contenu textuel, visuel, spatiotemporel, ainsi que l'engagement des utilisateurs et les messages publiés sur Twitter. Nous avons décrit en détail les étapes de construction de notre jeu de données. Pour montrer les caractéristiques distinctives entre les rumeurs et les informations factuelles, nous avons effectué de brèves analyses exploratoires à la fois sur les articles d'actualité et sur les messages correspondants de Twitter.

Par ailleurs, nous avons mené des expériences approfondies pour mesurer l'utilité des données de DAT@Z21 pour la tâche de classification de rumeurs. Pour cela, nous avons évalué plusieurs méthodes monomodales et multimodales récentes de l'état de l'art à l'aide des données disponibles dans DAT@Z21. Les résultats obtenus confirment que notre jeu de données renferme des caractéristiques pertinentes permettant la distinction des rumeurs et des informations factuelles. De plus, les méthodes multimodales qui exploitent plusieurs caractéristiques (textuelles, visuelles, etc.) réalisent les meilleures performances. C'est pourquoi un jeu de données contenant toutes ces caractéristiques, comme DAT@Z21, est nécessaire.

Pour permettre aux futures études de développer et tester de nouvelles techniques de détection et de classification des rumeurs, nous avons partagé DAT@Z21 avec la

communauté de recherche le jeu de données et toutes les explications nécessaires dans un GitLab à l'adresse https://git.msh-lse.fr/eric/dataz21.

Malgré les résultats expérimentaux satisfaisants, nous pouvons identifier plusieurs directions pour améliorer davantage la qualité de notre jeu de données. Premièrement, DAT@Z21 peut être étendu en incluant : (1) des articles d'actualité et des *tweets* dans d'autres langues, (2) d'autres sources d'information fiables, telles que d'autres organisations ou plateformes de vérification des faits évaluant la fiabilité et la partialité des sources d'information, afin de créer un vaste ensemble centralisé de labels de vérité terrain. En outre, étant donné que les rumeurs sont susceptibles d'être créées et diffusées par des comptes non-humains, tels que les robots sociaux [Shao et al., 2017], l'extension du jeu de données fournissant la vérité de terrain de ce type de messages peut être utile pour la classification des rumeurs.

Chapitre 7

Conclusion générale et perspectives

Sommaire

Pour conclure cette thèse, nous résumons tout d'abord dans la section 7.1 les contributions que nous avons présentées, puis nous présentons les principales perspectives de recherche ouvertes par ces travaux dans la section 7.2.

7.1 Synthèse des travaux

Dans ce manuscrit, nous nous sommes intéressés à la détection des rumeurs dans les sites de *microblogs* et avons présenté des solutions à ce problème, en nous focalisant plus spécifiquement sur des jeux de données issus de Twitter.

Premièrement, nous avons présenté et discuté l'état de l'art en relation avec les travaux présentés dans ce manuscrit. Cette revue de littérature est organisée en trois parties principales.

1. Nous résumons les approches pratiques proposées par les professionnels de l'information et de la communication.
2. Nous proposons une cartographie des solutions automatiques développées par le monde académique pour la détection des rumeurs en ligne.
3. Nous proposons une typologie des approches de vérification de la véracité d'une image numérique, notre volonté étant d'inclure les images attachées aux messages dans l'analyse de la véracité des messages (relatifs ou non à une fausse information).

L'état de l'art concernant les approches automatiques de détection de rumeurs et la véracité des images a fait l'objet d'une publication dans la conférence nationale EDA 2019 [Azri et al., 2019b].

Deuxièmement, nous avons proposé MONITOR (*Multimodal Fusion Framework to Assess Message Veracity in Social Networks*), une plateforme de fusion de plusieurs modalités des messages pour la vérification de leur véracité. MONITOR fonctionne en deux phases. La première est basée sur l'ingénierie des caractéristiques, dont le but est l'extraction de caractéristiques pour représenter fidèlement le contenu des messages. Plusieurs descripteurs sont développés à partir du texte et du contexte social des publications. Pour caractériser le contenu visuel des publications, notamment les images, nous avons proposé un ensemble de nouveaux descripteurs statistiques et visuels inspirés du domaine de l'évaluation de la qualité d'image (IQA). Ces métriques permettent d'estimer le taux de bruit et de quantifier la quantité de dégradation visuelle de tous types

dans une image. Dans la deuxième phase, les ensembles de caractéristiques des modalités sont concaténés pour former le vecteur de fusion comme une représentation multimodale finale du message. Plusieurs modèles supervisés d'apprentissage automatique sont appris pour la classification de la véracité du message. Ce travail a fait l'objet de deux publications dans les conférences internationales ADBIS 2021 [Azri et al., 2021b] et INTIS 2019 [Azri et al., 2019a].

Troisièmement, pour améliorer les performances de MONITOR, nous avons proposé l'utilisation de l'apprentissage ensembliste, un paradigme très peu exploité dans le domaine de la détection des rumeurs. Nous avons proposé cinq algorithmes (*soft voting, weighted average voting, stacking, blending* et *super learner ensemble*) qui utilisent les modèles appris dans MONITOR comme modèles de base. L'idée est que chaque modèle de base possède ses propres performances, qui diffèrent des autres. Autrement dit, les erreurs de prédiction faites par les modèles sont non-corrélées ou ont une faible corrélation, ce qui permet d'améliorer les performances de MONITOR. Cette contribution a fait l'objet d'un papier qui a été accepté en révision mineure en avril 2022 au journal Information Systems Frontiers chez Springer, suite à une invitation à étendre le papier d'ADBIS 2021.

Quatrièmement, nous avons proposé deepMONITOR, un modèle multimodal de bout en bout basé sur les techniques d'apprentissage profond, pour la classification de la véracité des rumeurs. deepMONITOR est composé essentiellement de réseaux CNN et RNN qui permettent d'extraire et de fusionner conjointement les informations de plusieurs modalités des messages. En plus du contenu textuel, la prise en compte du contenu visuel et des informations liées aux sentiments dans un contexte de détection de rumeur constitue la nouveauté de deepMONITOR. Cette contribution a fait l'objet d'une publication dans la conférence internationale ECML/PKDD 2021 [Azri et al., 2021a].

Cinquièmement, pour répondre au besoin en jeux de données pour la détection des rumeurs, nous avons conçu et mis à disposition DAT@Z21, un jeu de données multimodales complet. Ce jeu de données se distingue de l'existant d'une part, par la qualité de la vérité terrain collectée auprès des experts humains du site de vérification PolitiFact. D'autre part, DAT@Z21 présente un ensemble de données variées et riches en informations nécessaires pour la classification de rumeurs, comme le contenu textuel et linguistique, visuel, spatiotemporel, ainsi que les comportements des utilisateurs et leur engagement dans Twitter. Ce jeu de données volumineux a également été conçu en perspective d'explorer l'apprentissage profond, qui nécessite des données avec ces

propriétés. Nous avons de fait utilisé les données de DAT@Z21 pour l'évaluation de deepMONITOR.

Les expériences menées à l'aide de plusieurs jeux de données du monde réel, dont DAT@Z21, collectés sur Twitter, ont démontré la pertinence de nos propositions. Premièrement, en comparant les performances de MONITOR avec celles des méthodes exposées dans l'état de l'art, nous avons montré que les caractéristiques visuelles de l'image sont effectives dans la vérification de la véracité des messages. Nous avons également montré qu'avec la prise en compte de toutes les modalités des messages, MONITOR surpasse les autres méthodes. Deuxièmement, nous avons montré la pertinence de l'apprentissage ensembliste pour la tâche de détection des rumeurs. En effet, les cinq modèles ensemblistes sont plus performants que le meilleur modèle individuel de base (MONITOR). Troisièmement, nous avons montré que deepMONITOR réalise les meilleurs résultats par rapport aux techniques monomodales dans lesquelles les informations relatives au texte, aux images et aux sentiments sont utilisées séparément pour la classification des rumeurs. Nous avons également montré que deepMONITOR surpasse les techniques multimodales récentes de l'état de l'art. Quatrièmement, nous avons illustré à travers plusieurs analyses exploratoires les caractéristiques distinctives de DAT@Z21 entre les rumeurs et les informations factuelles. En outre, nous avons démontré l'utilité et la pertinence de notre jeu de données par rapport à plusieurs méthodes monomodales et multimodales récentes de l'état de l'art.

7.2 Perspectives

Au-delà des perspectives directes qui ont été proposées au fur et à mesure de la présentation des contributions, les travaux de recherche entamés durant cette thèse ouvrent plusieurs perspectives de recherche futures.

7.2.1 Comparaison entre MONITOR et deepMONITOR

Dans les chapitres 3 et 6 nous avons présenté nos deux modèles pour la détection des rumeurs dans les *microblogs*. MONITOR est un modèle basé sur l'ingénierie des caractéristiques et les algorithmes classiques d'apprentissage automatique. deepMONITOR est, lui, un modèle basé sur les techniques d'apprentissage profond. Les expériences

menées sur plusieurs jeux de données montrent que les deux modèles réalisent de bonnes performances de classification. Une piste de recherche intéressante serait de comparer les deux modèles pour étudier le compromis entre les performances de la classification, l'explicabilité, l'interprétabilité et la complexité du calcul.

Les questions relatives à l'explicabilité et à l'interprétabilité des modèles d'apprentissage jouent un rôle important, notamment dans le contexte de la détection de rumeurs. Appuyer les résultats de la classification par des éléments explicatifs pourraient être utiles à l'utilisateur pour démystifier la rumeur et empêcher sa propagation.

En outre, des études récentes [Schwartz et al., 2020; Strubell et al., 2019] commencent à s'intéresser aux questions relatives à l'impact écologique des modèles d'apprentissage. On parle de *Green AI* et de *Red AI*. Si nos modèles ont obtenu un gain notable de performance, quantifier leur coût environnemental (consommation d'énergie et empreinte carbone, par exemple) pourrait établir des points de comparaison plus pertinents que la performance pure.

7.2.2 Apprentissage non supervisé

Pour résoudre le problème de détection des rumeurs dans les *microblogs*, nous avons proposé dans cette thèse des modèles d'apprentissage supervisé. Une direction de recherche consiste à développer des modèles semi-supervisés ou non supervisés. Le principal avantage de l'exploitation de ces techniques réside dans le fait qu'elles peuvent apprendre à partir de données non étiquetées, ce qui résout le problème de la collecte des jeux de données avec des labels de qualité.

Un premier pas dans ce sens est proposé par Yang et al. [2019]. La véracité des informations et la crédibilité des utilisateurs sont traitées comme des variables aléatoires latentes dans un modèle de réseau bayésien. Le problème d'inférence est résolu au moyen d'une approche d'échantillonnage de Gibbs [Robert et al., 1999]. Cependant les auteurs comparent leur méthode avec d'autres algorithmes qui ne sont pas explicitement conçus pour la détection des rumeurs.

Nous imaginons trois types de modèles d'apprentissage non supervisés pour la détection des rumeurs.

1. **Analyse par grappes :** Dans l'analyse par grappes, les données peuvent être regroupées sur la base du principe de maximisation de la similarité intraclasse et de minimisation de la similarité interclasse, et elle peut générer des étiquettes de classe pour un groupe de données. Dans la détection des rumeurs, l'analyse en grappes peut être utilisée pour identifier, par exemple, l'homogénéité des groupes individuels de messages et d'auteurs.
2. **Analyse des données aberrantes :** L'analyse des données aberrantes est l'étude de la détection du comportement anormal d'objets. En utilisant l'apprentissage de la distribution statistique des données non labellisées des rumeurs, l'analyse des valeurs aberrantes peut découvrir des informations fausses et des auteurs suspects en se basant sur des mesures statistiques, des mesures de distance et des méthodes basées sur la densité.
3. **Analyse de la similarité sémantique :** L'analyse de similarité sémantique est utilisée pour détecter les contenus informatifs quasi-dupliqués [Ahmed, 2017]. En raison d'un manque de savoir-faire et d'imagination, les créateurs de rumeurs en ligne (appelés aussi pseudo-journalistes) réutilisent généralement le contenu d'actualités existantes. L'analyse de la similarité sémantique est donc un bon moyen de détecter un contenu entièrement ou partiellement copié et peut être utilisée pour la détection de potentielles rumeurs.

7.2.3 Outil de veille

Un aspect important du problème de détection des rumeurs est de mettre à la disposition des utilisateurs un outil de veille pour vérifier la véracité de nouvelles informations. Actuellement, il existe quelques services en ligne visant à aider les professionnels et les utilisateurs à effectuer des vérifications. TweetCred est un outil permettant d'évaluer la crédibilité du contenu sur Twitter [Gupta et al., 2014] . Le système fournit une note de crédibilité de 1 à 7 pour chaque *tweet* de la *timeline* Twitter (figure 2.5, page 26). Hoaxy est une plateforme pour la collecte, la détection des fausses informations en ligne et la vérification des faits correspondants [Shao et al., 2016]. Pour suivre la diffusion des fausses informations sur Twitter, Hoaxy collecte les *tweets* à partir de deux sources : (1) des sites web jugés de faible crédibilité qui publient souvent des affirmations inexactes et non vérifiées ; (2) des sites de vérification des faits, telles que Snopes, PolitiFact et Factcheck, qui vérifient régulièrement des publications sur les réseaux sociaux.

Développer un outil de veille en ligne pour la vérification des rumeurs doit répondre à

deux questions fondamentales. Premièrement, le volume massif, la haute dimensionnalité et la nature hétérogène des données en ligne posent des défis difficiles pour la construction d'un système de détection en temps réel, notamment en termes de stockage et de calcul des données. Deuxièmement, comment présenter les résultats de la vérification aux utilisateurs finaux d'une manière compréhensible ? En effet, un modèle de classification binaire est loin d'être suffisant pour discerner les caractéristiques des fausses informations en ligne. Nous pensons que la combinaison des résultats de la prédiction avec un module de visualisation des données, comme celui proposé par Zhao et al. [2014], peut être un outil efficace pour illustrer les différents aspects et schémas de distribution des informations sociales en ligne. Un système de visualisation interactif peut fournir diverses dimensions et vues des données, faciliter la supervision et la compréhension humaines, révéler des modèles et les comportements temporels des données et résumer les caractéristiques importantes de manière plus claire. Enfin, donner la possibilité aux utilisateurs de signaler de possible fausses informations constitue un aspect important pour améliorer les performances d'un outil de veille.

Bibliographie

Abd Warif, N. B., Wahab, A. W. A., Idris, M. Y. I., Ramli, R., Salleh, R., Shamshirband, S., and Choo, K.-K. R. (2016). Copy-move forgery detection : survey, challenges and future directions. *Journal of Network and Computer Applications*, 75 :259–278.

Abonizio, H. Q., de Morais, J. I., Tavares, G. M., and Barbon Junior, S. (2020). Language-independent fake news detection : English, portuguese, and spanish mutual features. *Future Internet*, 12(5) :87.

Afroz, S., Brennan, M., and Greenstadt, R. (2012). Detecting hoaxes, frauds, and deception in writing style online. In *2012 IEEE Symposium on Security and Privacy*, pages 461–475. IEEE.

Ahmed, H. (2017). *Detecting opinion spam and fake news using n-gram analysis and semantic similarity*. PhD thesis.

Ajao, O., Bhowmik, D., and Zargari, S. (2019). Sentiment aware fake news detection on online social networks. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2507–2511. IEEE.

Al-Ash, H. S., Putri, M. F., Mursanto, P., and Bustamam, A. (2019). Ensemble learning approach on indonesian fake news classification. In *2019 3rd International Conference on Informatics and Computational Sciences (ICICoS)*, pages 1–6. IEEE.

Allcott, H. and Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2) :211–36.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. (2015). Vqa : Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgments. *Organizational influence processes*, 58 :295–303.

Ashforth, B. E. and Mael, F. (1989). Social identity theory and the organization. *Academy of management review*, 14(1) :20–39.

Azri, A., Favre, C., Harbi, N., and Darmont, J. (2019a). Including images into message veracity assessment in social media. In *8th International Conference on Innovation and New Trends in Information Technology, Tangier, Morocco, 20-21 December 2019*.

Azri, A., Favre, C., Harbi, N., and Darmont, J. (2019b). Vers une analyse des rumeurs dans les réseaux sociaux basée sur la véracité des images : état de l'art. In Lemire, D. and Sautot, L., editors, *Business Intelligence Big Data, 15ème Edition de la conférence EDA, Montpellier, France, 3-4 octobre 2019*, volume B-15 of *RNTI*, pages 125–142. Éditions RNTI.

Azri, A., Favre, C., Harbi, N., Darmont, J., and Noûs, C. (2021a). Calling to cnn-lstm for rumor detection : A deep multi-channel model for message veracity classification in microblogs. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 497–513. Springer.

Azri, A., Favre, C., Harbi, N., Darmont, J., and Noûs, C. (2021b). Monitor : A multimodal fusion framework to assess message veracity in social networks. In *European Conference on Advances in Databases and Information Systems*, pages 73–87. Springer.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv :1409.0473*.

Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2) :157–166.

Bianchi, T. and Piva, A. (2012). Image forgery localization via block-grained analysis of jpeg artifacts. *IEEE Transactions on Information Forensics and Security*, 7(3) :1003–1017.

Boididou, C., Andreadou, K., Papadopoulos, S., Dang-Nguyen, D.-T., Boato, G., Riegler, M., Kompatsiaris, Y., et al. (2015). Verifying multimedia use at mediaeval 2015. *MediaEval*, 3(3) :7.

Boididou, C., Papadopoulos, S., Zampoglou, M., Apostolidis, L., Papadopoulou, O., and Kompatsiaris, Y. (2018). Detection and visualization of misleading content on twitter. *International Journal of Multimedia Information Retrieval*, 7(1) :71–86.

Cao, J., Guo, J., Li, X., Jin, Z., Guo, H., and Li, J. (2018). Automatic rumor detection on microblogs : A survey. *arXiv preprint arXiv :1807.03505*.

Castillo, C., Mendoza, M., and Poblete, B. (2011). Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684.

Chen, T., Li, X., Yin, H., and Zhang, J. (2018). Call attention to rumors : Deep attention based recurrent neural networks for early rumor detection. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 40–52. Springer.

Chen, W., Yeo, C. K., Lau, C. T., and Lee, B. S. (2016). Behavior deviation : An anomaly detection view of rumor preemption. In *2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pages 1–7. IEEE.

Cooke, N. A. (2017). Posttruth, truthiness, and alternative facts : Information behavior and critical information consumption for a new age. *The library quarterly*, 87(3) :211–221.

Cozzolino, D., Marra, F., Poggi, G., Sansone, C., and Verdoliva, L. (2017). Prnu-based forgery localization in a blind scenario. In *International Conference on Image Analysis and Processing*, pages 569–579. Springer.

Cui, L. and Lee, D. (2020). Coaid : Covid-19 healthcare misinformation dataset. *arXiv preprint arXiv :2006.00885*.

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

Dong, E., Du, H., and Gardner, L. (2020). An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*, 20(5) :533–534.

Flennerhag, S. (2017). mlens documentation.

Gentzkow, M., Shapiro, J. M., and Stone, D. F. (2015). Media bias in the marketplace : Theory. In *Handbook of media economics*, volume 1, pages 623–645. Elsevier.

Gers, F. A., Schraudolph, N. N., and Schmidhuber, J. (2002). Learning precise timing with lstm recurrent networks. *Journal of machine learning research*, 3(Aug) :115–143.

Giasemidis, G., Singleton, C., Agrafiotis, I., Nurse, J. R., Pilgrim, A., Willis, C., and Greetham, D. V. (2016). Determining the veracity of rumours on twitter. In *International Conference on Social Informatics*, pages 185–205. Springer.

Goljan, M., Fridrich, J., and Chen, M. (2010). Defending against fingerprint-copy attack in sensor-based camera identification. *IEEE Transactions on Information Forensics and Security*, 6(1) :227–236.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.

Gupta, A., Kumaraguru, P., Castillo, C., and Meier, P. (2014). Tweetcred : Real-time credibility assessment of content on twitter. In *International conference on social informatics*, pages 228–243. Springer.

Gupta, A., Lamba, H., Kumaraguru, P., and Joshi, A. (2013). Faking sandy : characterizing and identifying fake images on twitter during hurricane sandy. In *Proceedings of the 22nd international conference on World Wide Web*, pages 729–736.

Gupta, D. and Rani, R. (2020). Improving malware detection using big data and ensemble learning. *Computers & Electrical Engineering*, 86 :106729.

Gupta, M., Zhao, P., and Han, J. (2012). Evaluating event credibility on twitter. In *Proceedings of the 2012 SIAM International Conference on Data Mining*, pages 153–164. SIAM.

Hagberg, A., Swart, P., and S Chult, D. (2008). Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

Hamidian, S. and Diab, M. (2015). Rumor detection and classification for twitter data. In *SOTICS 2015*, pages 71–77.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8) :1735–1780.

Horne, B. and 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. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 759–766.

Huang, X., Wang, S., and Liu, G. (2018). Detecting double jpeg compression with same quantization matrix based on dense cnn feature. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 3813–3817. IEEE.

Huang, Y.-F. and Chen, P.-H. (2020). Fake news detection using an ensemble learning model based on self-adaptive harmony search algorithms. *Expert Systems with Applications*, 159 :113584.

Hutto, C. and Gilbert, E. (2014). Vader : A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Jin, Z., Cao, J., Guo, H., Zhang, Y., and Luo, J. (2017a). Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 795–816.

Jin, Z., Cao, J., Guo, H., Zhang, Y., Wang, Y., and Luo, J. (2017b). Detection and analysis of 2016 us presidential election related rumors on twitter. In *International conference on social computing, behavioral-cultural modeling and prediction and behavior representation in modeling and simulation*, pages 14–24. Springer.

Jin, Z., Cao, J., Jiang, Y.-G., and Zhang, Y. (2014). News credibility evaluation on microblog with a hierarchical propagation model. In *2014 IEEE international conference on data mining*, pages 230–239. IEEE.

Jin, Z., Cao, J., Zhang, Y., and Luo, J. (2016a). News verification by exploiting conflicting social viewpoints in microblogs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Jin, Z., Cao, J., Zhang, Y., and Zhang, Y. (2015). Mcg-ict at mediaeval 2015 : Verifying multimedia use with a two-level classification model. In *MediaEval*.

Jin, Z., Cao, J., Zhang, Y., Zhou, J., and Tian, Q. (2016b). Novel visual and statistical image features for microblogs news verification. *IEEE transactions on multimedia*, 19(3) :598–608.

Johnson, M. K. and Farid, H. (2006). Exposing digital forgeries through chromatic aberration. In *Proceedings of the 8th workshop on Multimedia and security*, pages 48–55.

Kaliyar, R. K., Goswami, A., and Narang, P. (2019). Multiclass fake news detection using ensemble machine learning. In *2019 IEEE 9th International Conference on Advanced Computing (IACC)*, pages 103–107. IEEE.

Kanske, P. and Kotz, S. A. (2010). Leipzig affective norms for german : A reliability study. *Behavior research methods*, 42(4) :987–991.

Karegowda, A. G., Manjunath, A., and Jayaram, M. (2010). Comparative study of attribute selection using gain ratio and correlation based feature selection. *International Journal of Information Technology and Knowledge Management*, 2(2) :271–277.

Kaur, S., Kumar, P., and Kumaraguru, P. (2020). Automating fake news detection system using multi-level voting model. *Soft Computing*, 24(12) :9049–9069.

Keras (Last accessed in 2021a). Repeatvector layer. https://keras.io/api/layers/reshaping_layers/repeat_vector/.

Keras (Last accessed in 2021b). Timedistributed layer. https://keras.io/api/layers/recurrent_layers/time_distributed/.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., and Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Kwon, S., Cha, M., and Jung, K. (2017). Rumor detection over varying time windows. *PloS one*, 12(1) :e0168344.

Kwon, S., Cha, M., Jung, K., Chen, W., and Wang, Y. (2013). Prominent features of rumor propagation in online social media. In *2013 IEEE 13th international conference on data mining*, pages 1103–1108. IEEE.

Lee, J., Wang, W., Harrou, F., and Sun, Y. (2020). Reliable solar irradiance prediction using ensemble learning-based models : A comparative study. *Energy Conversion and Management*, 208 :112582.

Li, J., Lu, W., Weng, J., Mao, Y., and Li, G. (2018). Double jpeg compression detection based on block statistics. *Multimedia Tools and Applications*, 77(24) :31895–31910.

Lin, D., Lv, Y., and Cao, D. (2015). Rumor diffusion purpose analysis from social attribute to social content. In *2015 International Conference on Asian Language Processing (IALP)*, pages 107–110. IEEE.

Lin, Z., Wang, R., Tang, X., and Shum, H.-Y. (2005). Detecting doctored images using camera response normality and consistency. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 1087–1092. IEEE.

Ma, B., Lin, D., and Cao, D. (2017). Content representation for microblog rumor detection. In *Advances in Computational Intelligence Systems*, pages 245–251. Springer.

Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B. J., Wong, K.-F., and Cha, M. (2016). Detecting rumors from microblogs with recurrent neural networks.

Ma, J., Gao, W., Wei, Z., Lu, Y., and Wong, K.-F. (2015). Detect rumors using time series of social context information on microblogging websites. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, pages 1751–1754.

Ma, J., Gao, W., and Wong, K.-F. (2018). Rumor detection on twitter with tree-structured recursive neural networks. Association for Computational Linguistics.

Maître, H. (2017). *From photon to pixel : the digital camera handbook*. John Wiley & Sons.

Martin, N. (2014). Information verification in the age of digital journalism. *Special Libraries Association (SLA)*.

Mihalcea, R. and Tarau, P. (2004). Textrank : Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Mitra, T. and Gilbert, E. (2015). Credbank : A large-scale social media corpus with associated credibility annotations. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 9, pages 258–267.

Mittal, A., Moorthy, A. K., and Bovik, A. C. (2011). Blind/referenceless image spatial quality evaluator. In *2011 conference record of the forty fifth asilomar conference on signals, systems and computers (ASILOMAR)*, pages 723–727. IEEE.

Mittal, A., Soundararajan, R., and Bovik, A. C. (2012). Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3) :209–212.

Morris, M. R., Counts, S., Roseway, A., Hoff, A., and Schwarz, J. (2012). Tweeting is believing ? understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*, pages 441–450.

Newman, M. L., Pennebaker, J. W., Berry, D. S., and Richards, J. M. (2003). Lying words : Predicting deception from linguistic styles. *Personality and social psychology bulletin*, 29(5) :665–675.

Ng, T.-T. and Chang, S.-F. (2004). A model for image splicing. In *2004 International Conference on Image Processing, 2004. ICIP'04.*, volume 2, pages 1169–1172. IEEE.

Nguyen, T. N., Li, C., and Niederée, C. (2017). On early-stage debunking rumors on twitter : Leveraging the wisdom of weak learners. In *International Conference on Social Informatics*, pages 141–158. Springer.

Nickerson, R. S. (1998). Confirmation bias : A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2) :175–220.

Nørregaard, J., Horne, B. D., and Adalı, S. (2019). Nela-gt-2018 : A large multi-labelled news dataset for the study of misinformation in news articles. In *Proceedings of the international AAAI conference on web and social media*, volume 13, pages 630–638.

Ott, M., Choi, Y., Cardie, C., and Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. *arXiv preprint arXiv :1107.4557*.

Pascanu, R., Mikolov, T., and Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pages 1310–1318. PMLR.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, 12 :2825–2830.

Pennebaker, J. W. and King, L. A. (1999). Linguistic styles : language use as an individual difference. *Journal of personality and social psychology*, 77(6) :1296.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove : Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Pérez-Rosas, V., Kleinberg, B., Lefevre, A., and Mihalcea, R. (2017). Automatic detection of fake news. *arXiv preprint arXiv :1708.07104*.

Pham, K., Kim, D., Park, S., and Choi, H. (2021). Ensemble learning-based classification models for slope stability analysis. *Catena*, 196 :104886.

Redondo, J., Fraga, I., Padrón, I., and Comesaña, M. (2007). The spanish adaptation of anew (affective norms for english words). *Behavior research methods*, 39(3) :600–605.

Robert, C. P., Casella, G., and Casella, G. (1999). *Monte Carlo statistical methods*, volume 2. Springer.

Ross, L., Ward, A., et al. (1996). Naive realism in everyday life : Implications for social conflict and misunderstanding. *Values and knowledge*, 103 :135.

Ruchansky, N., Seo, S., and Liu, Y. (2017). Csi : A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 797–806.

Sagi, O. and Rokach, L. (2018). Ensemble learning : A survey. *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery*, 8(4) :e1249.

Salem, F. K. A., Al Feel, R., Elbassuoni, S., Jaber, M., and Farah, M. (2019). Fa-kes : A fake news dataset around the syrian war. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 573–582.

Sangamnerkar, S., Srinivasan, R., Christhuraj, M., and Sukumaran, R. (2020). An ensemble technique to detect fabricated news article using machine learning and natural language processing techniques. In *2020 International Conference for Emerging Technology (INCET)*, pages 1–7. IEEE.

Schwartz, R., Dodge, J., Smith, N. A., and Etzioni, O. (2020). Green ai. *Communications of the ACM*, 63(12) :54–63.

Shahi, G. K. and Nandini, D. (2020). Fakecovid–a multilingual cross-domain fact check news dataset for covid-19. *arXiv preprint arXiv :2006.11343*.

Shao, C., Ciampaglia, G. L., Flammini, A., and Menczer, F. (2016). Hoaxy : A platform for tracking online misinformation. In *Proceedings of the 25th international conference companion on world wide web*, pages 745–750.

Shao, C., Ciampaglia, G. L., Varol, O., Flammini, A., and Menczer, F. (2017). The spread of fake news by social bots. *arXiv preprint arXiv :1707.07592*, 96 :104.

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., and Liu, H. (2020). Fakenewsnet : A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3) :171–188.

Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. (2017). Fake news detection on social media : A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1) :22–36.

Silverman, C. (2014). *Verification handbook : An ultimate guideline on digital age sourcing for emergency coverage*. European Journalism Centre.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv :1409.1556*.

Strubell, E., Ganesh, A., and McCallum, A. (2019). Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv :1906.02243*.

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.

Tacchini, E., Ballarin, G., Della Vedova, M. L., Moret, S., and de Alfaro, L. (2017). Some like it hoax : Automated fake news detection in social networks. *arXiv preprint arXiv :1704.07506*.

Trabelsi, A. and Zaiane, O. R. (2014). Mining contentious documents using an unsupervised topic model based approach. In *2014 IEEE International Conference on Data Mining*, pages 550–559. IEEE.

Van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super learner. *Statistical applications in genetics and molecular biology*, 6(1).

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Venkatanath, N., Praneeth, D., Bh, M. C., Channappayya, S. S., and Medasani, S. S. (2015). Blind image quality evaluation using perception based features. In *2015 Twenty First National Conference on Communications (NCC)*, pages 1–6. IEEE.

Vicario, M. D., Quattrociocchi, W., Scala, A., and Zollo, F. (2019). Polarization and fake news : Early warning of potential misinformation targets. *ACM Transactions on the Web (TWEB)*, 13(2) :1–22.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell : A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Wang, W. Y. (2017). " liar, liar pants on fire" : A new benchmark dataset for fake news detection. *arXiv preprint arXiv :1705.00648*.

Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., and Gao, J. (2018). Eann : Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, pages 849–857.

Wolpert, D. H. (1992). Stacked generalization. *Neural networks*, 5(2) :241–259.

Wu, K., Yang, S., and Zhu, K. Q. (2015). False rumors detection on sina weibo by propagation structures. In *2015 IEEE 31st international conference on data engineering*, pages 651–662. IEEE.

Wu, L. and Liu, H. (2018). Tracing fake-news footprints : Characterizing social media messages by how they propagate. In *Proceedings of the eleventh ACM international conference on Web Search and Data Mining*, pages 637–645.

Xiao, J. (2019). Svm and knn ensemble learning for traffic incident detection. *Physica A : Statistical Mechanics and its Applications*, 517 :29–35.

Yang, F., Liu, Y., Yu, X., and Yang, M. (2012). Automatic detection of rumor on sina weibo. In *Proceedings of the ACM SIGKDD workshop on mining data semantics*, pages 1–7.

Yang, S., Shu, K., Wang, S., Gu, R., Wu, F., and Liu, H. (2019). Unsupervised fake news detection on social media : A generative approach. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 5644–5651.

Zampoglou, M., Papadopoulos, S., and Kompatsiaris, Y. (2015). Detecting image splicing in the wild (web). In *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE.

Zhang, C. and Ma, Y. (2012). *Ensemble machine learning : methods and applications*. Springer.

Zhao, J., Cao, N., Wen, Z., Song, Y., Lin, Y.-R., and Collins, C. (2014). # fluxflow : Visual analysis of anomalous information spreading on social media. *IEEE transactions on visualization and computer graphics*, 20(12) :1773–1782.

Zhao, Z., Resnick, P., and Mei, Q. (2015). Enquiring minds : Early detection of rumors in social media from enquiry posts. In *Proceedings of the 24th international conference on world wide web*, pages 1395–1405.

Zheng, J., Liu, Y., Ren, J., Zhu, T., Yan, Y., and Yang, H. (2016). Fusion of block and keypoints based approaches for effective copy-move image forgery detection. *Multidimensional Systems and Signal Processing*, 27(4) :989–1005.

Zhou, X., Jain, A., Phoha, V. V., and Zafarani, R. (2019). Fake news early detection : An interdisciplinary study. *arXiv preprint arXiv :1904.11679*.

Zhou, X., Mulay, A., Ferrara, E., and Zafarani, R. (2020a). Recovery : A multimodal repository for covid-19 news credibility research. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 3205–3212.

Zhou, X., Wu, J., and Zafarani, R. (2020b). Safe : Similarity-aware multi-modal fake news detection. In *PAKDD*, pages 354–367. Springer.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.

Zhu, X. and Ghahramani, Z. (2002). Learning from labels and unlabeled data with label propagation.

Zhu, X., Ghahramani, Z., and Lafferty, J. D. (2003). Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International conference on Machine learning (ICML-03)*, pages 912–919.

Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., and Procter, R. (2018). Detection and resolution of rumours in social media : A survey. *ACM Computing Surveys (CSUR)*, 51(2) :32.