Détection d'anomalies dans les textes pour la veille

Yizhou Xu

▶ **To cite this version:**

Yizhou Xu. Détection d'anomalies dans les textes pour la veille. Informatique et langage [cs.CL]. Institut National des Langues et Civilisations Orientales- INALCO PARIS - LANGUES O', 2025. Français. ⟨NNT : 2025INAL0003⟩. ⟨tel-05106645⟩

HAL Id: tel-05106645

https://theses.hal.science/tel-05106645v1

Submitted on 11 Jun 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HAL Authorization

Institut National des Langues et Civilisations Orientales

École doctorale n°265

Langues, littératures et sociétés du monde

Équipe de Recherche Textes, Informatique, Multilinguisme (ER-TIM)

THÈSE

présentée par

Yizhou XU

soutenue le 22 janvier 2025
pour obtenir le grade de **Docteur de l'INALCO**
en Traitement Automatique des Langues

Détection d'anomalies dans les textes pour la veille

Thèse dirigée par :

Mme Frédérique SEGOND	Professeure associée, INALCO
Mme Kata GÁBOR	Maître de conférences, INALCO

RAPPORTEURS :

M. Eric GAUSSIER	Professeur des universités, Université Grenoble Alpes
M. Xavier TANNIER	Professeur des universités, Sorbonne Université

MEMBRES DU JURY :

M. Eric GAUSSIER	Professeur des universités, Université Grenoble Alpes
M. Xavier TANNIER	Professeur des universités, Sorbonne Université
Mme Elena CABRIO	Professeure des universités, Université Côte d'Azur
M. Cédric LOPEZ	Expert, Emvista
M. Mathieu ROCHE	Chercheur HDR, CIRAD
Mme Frédérique SEGOND	Professeure associée, INALCO
Mme Kata GÁBOR	Maître de conférences, INALCO

Laboratoire d'accueil

INaLCO, ER-TIM (EA 2520)
Équipe de Recherche Textes, Informatique, Multilinguisme
2 rue de Lille, 75007 Paris
http://www.er-tim.fr/

TABLE DES MATIÈRES

LISTE DES FIGURES

Liste des tableaux

Remerciement

Cette thèse s'inscrit dans une période marquée par de profonds bouleversements. Débutée en pleine crise sanitaire de la COVID-19, dans un climat d'incertitude et de contraintes imprévues, elle s'est déroulée dans un contexte soumis à des changements constants. Entre réorganisations successives de l'entreprise, transitions entre différentes équipes et déménagements fréquents, ce chemin sinueux a été jalonné de nombreux défis et ajustements, tant personnels que professionnels. Dans un tel contexte, le soutien et l'accompagnement de nombreuses personnes ont été bien plus qu'un simple appui : ils ont constitué un pilier essentiel sans lequel ce travail n'aurait pu aboutir. C'est donc avec une immense gratitude que je tiens à exprimer mes plus sincères remerciements.

Je dois d'abord mes remerciements les plus sincères à mes directrices de thèse, Frédérique et Kata, véritables Athéna de cette odyssée. Leur soutien constant, leur bienveillance et leurs conseils éclairés m'ont permis de naviguer dans les eaux tumultueuses de cette aventure. Depuis le premier jour de mon stage en master jusqu'à la veille de la soutenance de thèse, Frédérique m'a encadré avec rigueur et clairvoyance, guidant chaque étape, de l'élaboration du projet CIFRE à la rédaction finale. Son expertise, à la fois académique et industrielle, ainsi que sa vision élargie du domaine ont su éclairer mon chemin, dissiper mes doutes et m'ouvrir des perspectives que je n'aurais pu envisager seul. Kata, avec sa rigueur académique et sa connaissance approfondie du TAL, m'a accompagné et guidé pas à pas dans ce parcours, particulièrement au moment où le domaine a été bouleversé par l'avènement des LLMs. Son expertise et son soutien m'ont permis d'ajuster mon approche scientifique pour convertir ces bouleversements en leviers de progression.

Une gratitude particulière va à Leila, mon encadrante chez Bertin/ChapsVision, dont le soutien multiforme a été précieux tant pour la recherche que pour les défis du quotidien. Son sens pratique du monde industriel et sa rigueur scientifique ont durablement influencé cette thèse, même après son départ de l'entreprise. Sans ses conseils pour concilier exigences académiques et réalités industrielles, ce travail n'aurait pas atteint sa pleine mesure.

Je remercie sincèrement les membres du jury, Éric Gaussier, Xavier Tannier, Mathieu Roche, Elena Cabrio et Cédric Lopez, pour le temps consacré à l'évaluation de mes travaux, ainsi que pour la richesse de nos échanges qui ont considérablement enrichi ce travail. Ma reconnaissance s'étend également à Elena et Pierre, membres de mon comité de suivi, pour leur soutien continu tout au long de ce parcours.

L'équipe ERTIM de l'INaLCO mérite aussi toute ma gratitude, en particulier Damien, Mathieu et Ilaine, pour leur aide précieuse dans les aspects administratifs et bien au-delà. Mes remerciements vont également à mes collègues doctorants, Boyu, Jinyuan, Manying, Johanna et Kevin, pour nos échanges réguliers, qui m'ont permis de rester connecté à la vie académique malgré mon ancrage en entreprise.

Je souhaite également remercier les membres de l'école doctorale, en particulier Mme Barbora Chavel, pour leur aide précieuse dans mes démarches administratives et dans l'organisation de la soutenance.

Je tiens à exprimer ma profonde gratitude à ChapsVision pour m'avoir offert l'opportunité de réaliser cette thèse CIFRE dans des conditions idéales, et pour la confiance accordée tout au long de ce projet. Mes remerciements s'adressent particulièrement aux équipes avec lesquelles j'ai eu la chance de collaborer au fil des années. Le groupe MédiaCentrique, à Paris, m'a accueilli dès mon stage de master et au début de ma thèse ; je garde un souvenir reconnaissant de cette première immersion, en particulier grâce à Paule, Franck, Guillaume et Florian. Le groupe AMI m'a ensuite chaleureusement accueilli à Montpellier. Je remercie ses membres pour leur accompagnement constant au fil des années, et j'adresse une pensée particulière à Sébastien Marinier, Sébastien Seron, Virginie, Nathan, Rachid, Jacques, Roman, Jérôme, Aurélien et Clément pour leur aide et leur bienveillance. Je remercie également les membres de ChapsLab et du groupe NLP pour leur collaboration et leur soutien, en particulier Aurélie, Zhen et Saïd, pour leur disponibilité et leur aide précieuse durant les périodes les plus instables de l'entreprise.

Je remercie l'ANRT pour son soutien financier dans le cadre du dispositif CIFRE, qui m'a permis de mener cette recherche dans des conditions favorables à l'interface entre le monde académique et industriel.

Je tiens à exprimer toute ma gratitude à ma famille et mes amis pour leur présence, leur écoute et leur soutien inconditionnel, même à distance. Leur confiance m'a accompagné dans les moments d'incertitude comme dans les instants de joie.

Enfin, mes remerciements les plus profonds vont à Chunyang, compagnon de vie et de route depuis plus de quinze ans, la moitié de mon existence. Ensemble, nous avons traversé les saisons et les frontières, de la Chine à l'Europe, en partageant les chemins de Lyon, Paris, Montpellier et Genève. Face aux défis, aux changements et aux silences du quotidien, ta présence constante m'a apporté force, stabilité et réconfort. Cette thèse porte aussi ton empreinte.

Au terme de ce parcours, je mesure la chance d'avoir été entouré de tant de personnes inspirantes et bienveillantes. Que ces remerciements soient un témoignage de ma gratitude pour celles et ceux qui ont rendu cette aventure possible.

Introduction générale

Enjeux

Au cours des dernières années, le monde a traversé une période de turbulences sans précédent. La fin de 2020 à 2024, période pendant laquelle cette thèse a été rédigée et ses recherches menées, a été marquée par des crises globales et locales telles que la pandémie de COVID-19 et la guerre russo-ukrainienne. Ces événements ont pris de court les gouvernements, les entreprises et tant d'autres organisations, perturbant profondément leurs environnements de développement et de fonctionnement. L'émergence soudaine de ces incidents majeurs n'a pas seulement modifié les dynamiques mondiales, mais a également révélé les vulnérabilités des plans stratégiques existants, entraînant des conséquences graves telles que des récessions économiques, des faillites d'entreprises et des dysfonctionnements gouvernementaux.

Cependant, malgré l'apparente imprévisibilité de ces crises, elles ne surgissent souvent pas sans avertissement. Des signaux faibles précèdent généralement ces événements transformateurs, fournissant des indices subtils de changements potentiels. Dans le domaine des données textuelles, ces signaux faibles se manifestent souvent sous la forme de textes anormaux, c'est-à-dire des informations qui dévient des schémas ou thèmes attendus au sein d'un flux d'information donné. Identifier de telles anomalies est crucial car elles peuvent servir d'alertes précoces, permettant la mise en place de mesures proactives en matière de gestion des risques et de planification stratégique.

La détection d'anomalies textuelles devient donc un outil essentiel pour identifier ces signaux faibles. En analysant les déviations dans les données textuelles, les organisations peuvent obtenir des informations sur les risques et opportunités émergents, améliorant ainsi leur préparation et leur capacité à répondre aux changements imprévus. Par exemple, dans le contexte de la veille économique, la capacité à détecter des anomalies dans les rapports de marché ou les conversations sur les réseaux sociaux peut aider les organisations à anticiper les perturbations du marché ou les changements dans le sentiment des consommateurs. À mesure que l'environnement mondial devient de plus en plus volatile, la capacité à reconnaître et à agir sur ces indicateurs précoces est cruciale pour maintenir la résilience et l'avantage stratégique.

Contexte de la recherche

Contexte industriel Ce travail de thèse est réalisé au sein de l'entreprise Bertin IT (actuellement ChapsVision CyberGov) dans le cadre d'une convention CIFRE. Bertin IT est une entreprise spécialisée dans le développement de solutions logicielles pour la cybersécurité, l'intelligence stratégique et le traitement automatique du langage naturel. Le principal objectif de ce travail de recherche est d'enrichir les fonctionnalités de deux plateformes clés de Bertin IT : MediaCentric et AMI Enterprise

Intelligence (AMI EI).

MediaCentric est une plateforme d'investigation en profondeur qui permet l'acquisition multi-sources (web, médias, réseaux sociaux) en temps réel et l'analyse de contenus multimédias et multilingues. Elle est principalement utilisée par les agences gouvernementales pour la détection de signaux faibles et l'anticipation de menaces. La détection d'anomalies textuelles y joue un rôle crucial pour alerter rapidement les analystes sur des informations divergentes ou inhabituelles qui pourraient indiquer des événements émergents ou des crises.

AMI EI, quant à elle, est une solution de veille stratégique qui aide les entreprises à exploiter les données massives (*Big Data*) pour anticiper les évolutions de leur environnement concurrentiel, technologique ou législatif, et pour identifier de nouvelles opportunités de développement. La détection d'anomalies textuelles dans ce contexte permet de repérer des tendances naissantes, des innovations technologiques ou des changements réglementaires susceptibles d'influencer le positionnement stratégique de l'entreprise.

Compte tenu de ce contexte, cette étude se concentre sur la détection d'anomalies textuelles appliquée aux systèmes de veille. L'objectif est de développer et d'adapter des méthodes de détection capables d'identifier des anomalies dans des flux de données textuelles variés. Cette recherche vise à améliorer la capacité des plateformes MediaCentric et AMI EI à repérer des informations déviantes, afin de soutenir les processus de veille stratégique et d'anticipation des évolutions du marché.

Contexte académique La détection d'anomalies, largement étudiée depuis des décennies dans le domaine de la fouille de données, a abouti au développement de nombreux algorithmes d'apprentissage automatique. Cependant, les recherches actuelles dans ce domaine se concentrent principalement sur des types de données structurées, telles que les données tabulaires, les séries temporelles ou les images. En ce qui concerne les données textuelles non structurées, peu d'études ont été menées pour adapter ces méthodes à la spécificité du texte. L'un des objectifs de cette thèse est donc d'explorer comment étendre les méthodes de fouille de données, en particulier celles récemment proposées, afin de mieux prendre en compte les caractéristiques des données textuelles.

Se situant à l'intersection de la fouille de données et du traitement automatique des langues naturelles (TALN), la détection d'anomalies textuelles a progressivement attiré l'attention de la communauté TALN ces dernières années. Plusieurs travaux de recherche ont introduit des techniques de TALN, notamment les modèles préentraînés, pour améliorer la détection d'anomalies dans les textes. Dans cette continuité, un autre objectif de cette thèse est d'explorer l'utilisation des grands modèles de langue (*Large Language Models*, LLMs) pour la détection d'anomalies textuelles, en mettant l'accent sur leur capacité à capturer des nuances contextuelles et sémantiques dans les données textuelles non structurées.

Ces axes de recherche visent à combler les lacunes existantes dans l'application des méthodes de détection d'anomalies au texte, tout en tirant parti des avancées récentes en apprentissage automatique et en TALN.

Organisation du manuscrit

Pour répondre aux objectifs de cette recherche, le manuscrit est structuré en trois parties principales :

- ✏ **Partie I : État de l'art**

  La première partie explore les méthodes de détection d'anomalies existantes dans les domaines de la fouille de données et du traitement automatique des langues. Elle met en lumière les défis spécifiques posés par les données textuelles non structurées.

- ✏ **Partie II : Application des méthodes de fouille de données appliquées aux textes**

  La deuxième partie présente l'adaptation et l'extension des techniques de fouille de données pour la détection d'anomalies textuelles. Il examine comment ces méthodes peuvent être appliquées aux flux de données textuelles pour repérer des informations atypiques.

- ✏ **Partie III : Utilisation des grands modèles de langue pour la détection d'anomalies**

  La dernière partie explore l'intégration des LLMs dans la détection d'anomalies textuelles, en évaluant leur capacité à capter des informations contextuelles et sémantiques. Elle aborde également les défis liés à leur déploiement dans des systèmes de veille.

À travers les différentes parties de ce manuscrit, nous cherchons à répondre aux questions de recherche suivantes :

— Quels sont les principaux facteurs qui influencent l'efficacité des méthodes de fouille de données dans la détection d'anomalies pour les données textuelles non structurées ?

— Comment les LLMs se comparent-ils aux méthodes traditionnelles de fouille de données dans la détection d'anomalies textuelles ? Dominent-t-ils dans ce domaine comme dans d'autres ?

— Comment les contraintes pratiques, telles que les ressources computationnelles, le temps de traitement et la disponibilité des données, influencent-elles le choix des techniques de détection d'anomalies textuelles ?

— Quelles sont les perspectives de recherche pour améliorer la détection d'anomalies textuelles, notamment avec l'intégration des LLMs ?

Première partie

État de l’art

Introduction de la première partie

Cette première partie de la thèse est consacrée à l'établissement d'un état de l'art détaillé sur la détection d'anomalies, avec une attention particulière portée aux anomalies dans les textes. Notre objectif est de dresser un panorama des recherches et développements récents dans ce domaine, en identifiant les tendances actuelles, les méthodes prédominantes et les lacunes éventuelles. Cette revue permet de situer la détection d'anomalies textuelles dans un contexte scientifique plus large, offrant ainsi une base théorique et pratique nécessaire pour notre recherche.

* * *

Le Chapitre 1 offre un aperçu général de la recherche en détection d'anomalies. Nous commencerons par un examen des bases théoriques et pratiques du domaine. Ensuite, nous nous pencherons sur les différentes approches et techniques utilisées dans la détection d'anomalies, suivies d'une discussion sur les métriques d'évaluation prédominantes pour les tâches et les systèmes de détection d'anomalies. Ce chapitre se terminera par une revue des applications dans divers domaines, en mettant en lumière les défis spécifiques posés par différents types de données.

Dans le Chapitre 2, nous nous concentrons sur la détection d'anomalies dans les données textuelles, objectif poursuivi dans ce travail de thèse. Nous commencerons par les concepts clés des anomalies textuelles, ainsi que des ressources linguistiques existantes. Ensuite, nous passerons en revue les différentes approches méthodologiques pour la détection d'anomalies textuelles. Enfin, nous aborderons spécifiquement la détection d'anomalies dans le contexte de la veille, en définissant la portée de notre recherche et ses applications potentielles.

* * *

Ainsi, à travers ces chapitres, nous positionnons la détection d'anomalies textuelles dans un cadre élargi, reliant les connaissances théoriques de la détection d'anomalies à leurs applications spécifiques dans le domaine des textes. Cette revue de la littérature vise à répondre aux questions de recherche suivantes :

— Comment les anomalies sont-elles définies dans la littérature, et quelles sont les caractéristiques principales qui les distinguent des données normales ?
— Quel est le processus général de détection d'anomalies, et quelles sont les étapes typiques impliquées ?
— Quelles sont les applications et les cas d'utilisation communs de la détection d'anomalies dans divers domaines ?
— Quelles sont les approches et les techniques les plus couramment utilisées pour la détection des anomalies ? Comment les approches d'apprentissage automatique (non supervisées, semi-supervisées et faiblement supervisées) diffèrent-

elles dans leur application à la détection d'anomalies ? Quels sont les apports des techniques d'apprentissage profond à la détection d'anomalies par rapport aux méthodes traditionnelles ?

— Quelles sont les tendances actuelles dans la recherche en détection d'anomalies, et quelles lacunes ou questions ouvertes subsistent ?

— Quelles sont les métriques couramment utilisées pour évaluer la performance des modèles de détection d'anomalies ? Comment différentes métriques d'évaluation impactent-elles le jugement porté sur les modèles ?

— Quels phénomènes linguistiques peuvent être abordés dans le cadre de la détection d'anomalies ?

— Quelles sont les caractéristiques clés et les types d'anomalies textuelles ?

— Quels sont les défis dans la création de jeux de données de référence pour la détection d'anomalies, et comment sont-ils généralement abordés ?

— Comment la détection d'anomalies textuelles évolue-t-elle avec l'avènement des grands modèles de langue ?

— Quels sont les domaines d'application potentiels pour la détection d'anomalies dans les textes, en particulier dans le contexte de la veille, et quels défis spécifiques présentent-ils ?

En abordant ces questions, nous visons à établir les axes de notre propre recherche. Cette approche mettra en lumière les avancées actuelles et révélera les domaines nécessitant une exploration plus approfondie, fournissant ainsi une base solide pour notre investigation.

CHAPITRE 1

DÉTECTION D'ANOMALIES

Sommaire

1.1 Introduction

La détection d'anomalies est un domaine de recherche dynamique et très actif, qui se situe à l'intersection de plusieurs disciplines scientifiques et technologiques, notamment l'ingénierie, la statistique, l'apprentissage automatique et la fouille de données. Il s'agit d'identifier des patterns ou des points de données qui s'écartent de manière significative du comportement attendu, tel que défini par la majorité des données. Ces anomalies sont souvent indicatives d'erreurs, de changements significatifs ou d'activités suspectes, rendant leur détection cruciale dans de nombreux secteurs. En reconnaissant ces déviations, la détection d'anomalies fournit des informations critiques qui s'avèrent précieuses dans divers contextes.

L'importance de la détection d'anomalies est soulignée par son application répandue dans de nombreux domaines clés :

— **Finance** : Identifier les transactions frauduleuses [Ahmed et al., 2016a; Huang et al., 2018; Moschini et al., 2021; Hilal et al., 2022] et les mouvements de marché anormaux [Luo et al., 2008; Golmohammadi and Zaiane, 2015; Ahmed et al., 2017; Weber et al., 2019; Cheong et al., 2021], pour éviter les pertes financières et renforcer la sécurité des systèmes financiers.

— **Sécurité informatique** : Reconnaître les signes d'intrusion dans les systèmes ou de cyber-attaques [Garcia-Teodoro et al., 2009; Xu et al., 2009; Gogoi et al., 2011; Ten et al., 2011; Pascoal et al., 2012; Vartouni et al., 2018; Meng et al., 2019], protégeant ainsi les informations sensibles et les infrastructures contre les activités malveillantes.

— **Veille stratégique** : Surveiller les médias et les réseaux sociaux pour détecter les tendances émergentes, les nouvelles concurrences et les changements dans l'opinion publique, ce qui aide à la prise de décisions stratégiques [Saranya et al., 2014; Takahashi et al., 2014; Yu et al., 2016; Kim and Lee, 2017; Sufi, 2022].

— **Santé et médical** : Détecter des schémas inhabituels dans les données des patients [Hauskrecht et al., 2013; Salem et al., 2013; Salmon et al., 2016; Su et al., 2019; Liu et al., 2020b; Gupta et al., 2021], ce qui facilite le diagnostic précoce et la personnalisation des traitements.

— **Industrie** : Dans l'industrie, notamment avec l'avènement de l'internet des objets, la détection d'anomalies devient cruciale pour la surveillance de l'état des machines connectées et la prévision de pannes [Fahim and Sillitti, 2019; Janjua et al., 2019; Cook et al., 2020; Al-amri et al., 2021; Ullah and Mahmoud, 2021], ce qui contribue à la maintenance préventive et à l'amélioration des processus de production.

Ce premier chapitre de l'état de l'art a pour objectif de fournir une vision globale de la recherche en détection d'anomalies. Nous commencerons avec la présentation des fondements théoriques et pratiques de ce domaine. Ensuite, nous procéderons à l'analyse des approches et techniques courantes sous différents angles, et à la discussion des critères d'évaluation prédominants pour les tâches et systèmes de détection d'anomalies. En fin, nous aborderons les applications de la détection d'anomalies dans divers domaines, en mettant particulièrement l'accent sur les défis spécifiques posés par différents types de données.

1.2 Cadre théorique et pratique

1.2.1 Anomalie

Dans de nombreux domaines du monde réel, des données sont constamment générées à travers divers processus naturels et humains. Que ce soit dans les systèmes environnementaux, socio-économiques, mécaniques ou biologiques, chaque processus suit généralement un certain modèle ou comportement prévisible, formant ainsi une « norme » pour les données produites. Cette référence normative reflète l'état normal et fonctionnel du système ou du processus en question.

Cependant, au sein de ce flux continu de données, il se produit parfois des points de données qui ne sont pas conformes au comportement attendu. Ces données aberrantes ne sont pas de simples coïncidences ; elles signifient souvent que les processus sous-jacents ont subi des modifications ou fonctionnent dans des conditions nouvelles

et inattendues. Ainsi, elles peuvent servir d'indicateurs d'événements significatifs, tels que des changements environnementaux, des tendances économiques imprévues, des défaillances dans les systèmes mécaniques, ou même de nouvelles dynamiques sociales, suscitant ainsi un grand intérêt dans divers domaines de recherche.

Dans la littérature scientifique, la terminologie utilisée pour désigner ces déviations significatives varie selon le contexte. Des termes comme « nouveauté (*novelty*) » [Schölkopf et al., 1999; Yang et al., 2002; Markou and Singh, 2003a,b; Hoffmann, 2007; Pimentel et al., 2014; Tack et al., 2020], « déviation (*deviation* ou *deviant*) » [Arning et al., 1996; Palpanas et al., 2003; Kamaruddin et al., 2015; Toth and Chawla, 2018; Foorthuis, 2021], « exception (*exception*) » [Knorr, 2002; Suzuki et al., 2003; Luo et al., 2008; Ruiz et al., 2015; Riahi and Schulte, 2018], « aberration (*aberration*) » [Hutwagner et al., 2005; Jackson et al., 2007; Salmon et al., 2016; Yuan et al., 2019b], « surprise (*surprise*) » [Bayarri and Morales, 2003; Itti and Baldi, 2009; Borne and Vedachalam, 2012; Gutflaish et al., 2019], « évènement rare (*rare event*) » [Cheon et al., 2009; Straub et al., 2016; Janjua et al., 2019; Theofilatos et al., 2019] et « point de changement (*change point*) » [Takeuchi and Yamanishi, 2006; Zhou et al., 2016; Aminikhanghahi and Cook, 2017; Fearnhead and Rigaill, 2019] sont couramment employés, chacun portant des nuances qui correspondent à des scénarios particuliers dans différents domaines. En fouille de données, où le sujet est étudié intensivement depuis des décennies, les termes « anomalie (*anomaly*) » [Chandola et al., 2009; Ahmed et al., 2016a; Chalapathy et al., 2019; Pang et al., 2021b; Ait-Saada and Nadif, 2023; Bejan et al., 2023; Samariya and Thakkar, 2023; Xu et al., 2023b; Su et al., 2024] et « valeur aberrante (*outlier*) » [Aggarwal, 2017a; Kannan et al., 2017; Zhuang et al., 2017; Aggarwal and Reddy, 2018; Domingues et al., 2018; Barrett et al., 2019; Boukerche et al., 2021; Xu et al., 2023a] sont largement utilisés de manière interchangeable. Pour maintenir une terminologie cohérente au sein de ce manuscrit de thèse, nous adoptons dorénavant le terme « anomalie ».

Cette diversité terminologique reflète la nature intrinsèquement floue et contextuelle des anomalies. Tout comme il n'existe pas de consensus dans la littérature sur l'usage de ces termes, il n'y a pas non plus de définition universellement acceptée du terme « anomalie » lui-même. Par leur nature même, les anomalies sont inconnues, inattendues et imprévisibles, ce qui complique toute tentative de les prédéfinir de manière uniforme. En général, il convient de les définir selon le domaine d'application et de les délimiter en fonction du contexte spécifique.

Pourtant, une définition fréquemment citée et largement adoptée, proposée par Hawkins [1980], peut constituer un point de départ pour notre étude :

Définition (Anomalie)**.** Une anomalie est une observation qui s'écarte tellement des autres observations qu'elle suscite des soupçons qu'elle a été générée par un mécanisme différent.

Cette définition encapsule l'essence des anomalies comme indicateurs de changements ou de perturbations significatifs. La détection de ces anomalies est cruciale, car elle sert de système d'alerte précoce pour des problèmes potentiels ou des évolutions importantes. Dans des systèmes industriels, par exemple, une perturbation dans les données de performance d'un équipement peut être le premier signe d'une défaillance imminente. Sa détection permet alors de planifier une maintenance proactive et d'éviter des interruptions de production coûteuses. De même, dans le domaine

de la cybersécurité, des changements inhabituels dans le trafic réseau peuvent indiquer une tentative d'intrusion, et leur détection permet d'agir avant que des dommages ne soient causés. Ainsi, ce sont les changements et perturbations signalés par les anomalies qui nous offrent une opportunité d'identifier et d'intervenir sur des problèmes cachés ou émergents. Ce n'est pas simplement le fait que l'anomalie s'écarte de la norme qui est important, mais plutôt qu'elle constitue souvent le signe d'un dysfonctionnement majeur ou d'une modification sous-jacente du système.

De plus, en définissant les anomalies simplement comme des points de données qui s'écartent de manière significative de la majorité dans un ensemble de données, cette définition établit des principes fondamentaux pour la pratique de la détection d'anomalies.

1.2.2 Détection d'anomalies

Selon la définition proposée par Hawkins [1980], la détection d'anomalies est un processus analytique visant à identifier des observations qui dévient significativement de la majorité des données dans un ensemble donné, comme illustré dans la Figure 1.1. En pratique, le problème de la détection d'anomalies peut être formulé comme suit :

Données d'entrée Soit un ensemble de données $\mathcal{X}$ composé de n observations, où chaque observation $x_i \in \mathbb{R}^d$ est représentée par un vecteur à d dimensions :

$$\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$$

Cet ensemble de données $\mathcal{X}$ est composé de deux sous-ensembles : le sous-ensemble d'anomalies $\mathcal{A}$ et le sous-ensemble de données normales $\mathcal{N}$, avec $|\mathcal{A}| \ll |\mathcal{N}|$. Cela correspond au scénario typique de la détection d'anomalies, où les anomalies sont extrêmement rares par rapport aux données normales.

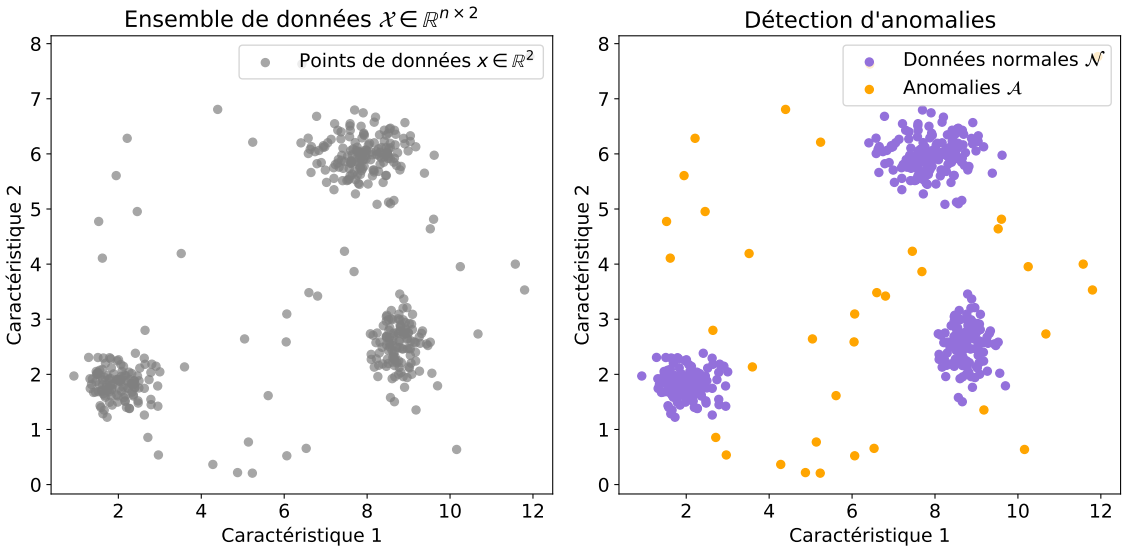
Objectif L'objectif de la détection d'anomalies est de construire un modèle qui peut soit :

1. **Fournir des étiquettes binaires** : Indiquer si chaque instance est normale (`0`) ou anormale (`1`).
2. **Produire des scores d'anomalie** : Indiquer le degré de déviation de chaque instance.

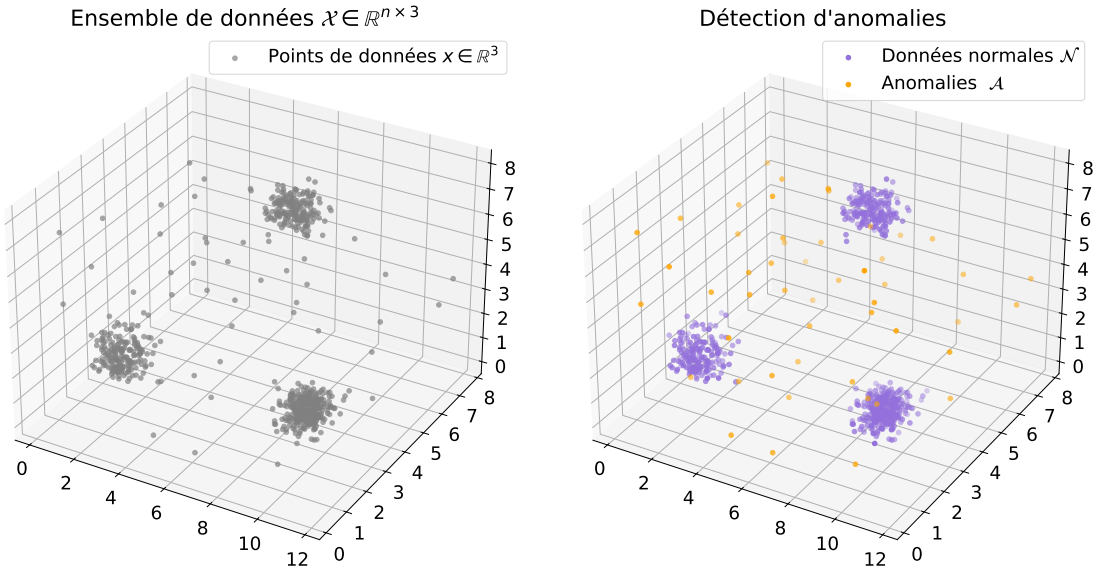
✏ **Étiquettes binaires** Dans l'approche d'étiquetage binaire, le modèle attribue une étiquette $y_i \in \{0, 1\}$ à chaque observation x_i, où :

$$y_i = \begin{cases} 1 & \text{si } x_i \in \hat{\mathcal{A}} \\ 0 & \text{si } x_i \in \hat{\mathcal{N}} \end{cases}$$

Ici, $\hat{\mathcal{A}}$ représente l'ensemble des anomalies prédites, et $\hat{\mathcal{N}}$ représente l'ensemble des instances normales prédites. Le but est de concevoir un modèle qui attribue exactement ces étiquettes binaires en fonction des caractéristiques des données concernées.



(a) Détection d'anomalies dans un espace 2D.



(b) Détection d'anomalies dans un espace 3D.

FIGURE 1.1 – Illustration de la détection d'anomalies dans les espaces 2D et 3D. Cette figure présente la détection d'anomalies dans les ensembles de données $\mathcal{X} \in \mathbb{R}^{n \times 2}$ et $\mathcal{X} \in \mathbb{R}^{n \times 3}$. Les points identifiés comme normaux ($\mathcal{N}$) sont en violet, tandis que les points détectés comme des anomalies ($\mathcal{A}$) sont en orange.

✏ **Scores d'anomalie** Dans l'approche de scoring des anomalies, le modèle attribue un score à valeur réelle $\psi(x_i) \in \mathbb{R}$ à chaque observation x_i, indiquant son degré de déviation. Idéalement, la fonction $\psi : \mathcal{X} \mapsto \mathbb{R}$ devrait être conçue de telle sorte que pour toute anomalie véritable $x_i \in \mathcal{A}$ et tout objet de données normal véritable $x_j \in \mathcal{N}$, la fonction satisfasse :

$$\psi(x_i) > \psi(x_j)$$

De cette façon, les scores les plus élevés indiquent une plus grande probabilité d'être des anomalies. Pour déterminer quelles instances sont des anomalies, un seuil τ est prédéfini. Les observations pour lesquelles $\psi(x_i) > \tau$ sont étiquetées comme anomalies :

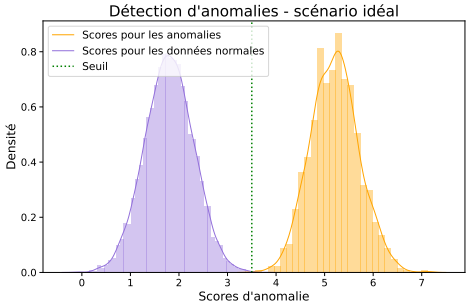
$$\hat{\mathcal{A}} = \{x_i \in \mathcal{X} \mid \psi(x_i) > \tau\}$$

Il est ainsi possible de prendre des décisions de manière flexible en fonction des exigences spécifiques de l'application.

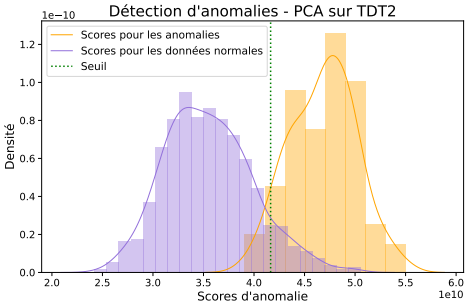
Procédure L'approche d'étiquetage binaire offre une réponse directe et claire, ce qui la rend particulièrement adaptée aux applications nécessitant une prise de décision rapide, telles que les systèmes de sécurité et les systèmes d'alerte précoce. Cependant, il manque de flexibilité et d'interprétabilité par rapport à l'approche basée sur les scores d'anomalie. Les scores d'anomalie permettent non seulement de prendre des décisions binaires en appliquant des seuils, mais aussi de quantifier la sévérité des anomalies, ce qui renforce le processus de prise de décision en fournissant des aperçus précis sur la nature et la magnitude de chaque déviation.

L'approche à base de scores d'anomalie est ainsi privilégiée dans la plupart des systèmes de détection d'anomalies du monde réel. Par conséquent, une solution typique au problème de la détection d'anomalies implique les étapes suivantes :

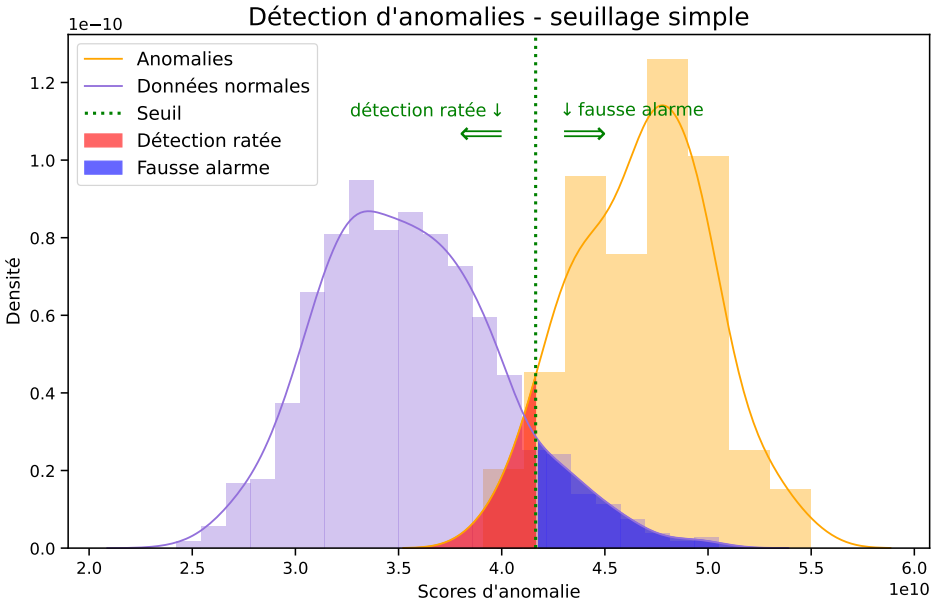
1. **Modélisation du comportement normal** : Cette première étape consiste à construire un modèle $\mathcal{M}$, qui capture le comportement normal de l'ensemble de données. Selon la disponibilité des étiquettes de vérité terrain à cette phase, ce modèle peut être développé sous divers paradigmes d'apprentissage, tels que l'apprentissage non supervisé, semi-supervisé ou faiblement supervisé.
2. **Calcul des scores d'anomalie** : Avec le modèle $\mathcal{M}$ en place, l'étape suivante consiste à évaluer chaque observation x_i pour déterminer sa déviation par rapport à ce qui est défini comme normal. Cette déviation est quantifiée par un score d'anomalie $\psi(x_i)$, qui est calculé sur la base de différentes théories sous-jacentes appropriées aux données.
3. **Seuil de décision** : La dernière étape consiste à fixer un seuil τ qui détermine quels scores indiquent un comportement anormal. Ce seuil peut être soit une valeur statique, basée sur des connaissances préalables et des données historiques, soit une valeur dynamique, mise à jour en réponse à l'évolution des caractéristiques de données.



(a) Scénario idéal



(b) Détection d'anomalies en conditions réelles



(c) Seuillage de la détection d'anomalies

FIGURE 1.2 – Illustration du seuillage de la détection d'anomalies. (a) Scénario idéal. Présenter deux distributions sans chevauchement où les scores d'anomalie pour les observations anormales ($\mathcal{A}$) et normales ($\mathcal{N}$) sont clairement séparés, facilitant l'établissement d'un seuil efficace à $\tau = 0.35$. (b) Détection d'anomalies en conditions réelles. Illustrer les distributions chevauchantes des scores d'anomalie produits en contexte pratique par l'algorithme PCA sur le corpus TDT2. (c) Seuillage. Montre l'impact de l'emplacement du seuil sur la prise de décision, en mettant en lumière les régions correspondant aux erreurs de détection. La zone de chevauchement à gauche du seuil (en rouge) correspond au taux de ratés, représentant la proportion d'anomalies réelles mal classifiées comme normales. Inversement, la zone de chevauchement à droite du seuil (en bleu) correspond au taux de fausses alarmes, indiquant la proportion de données normales incorrectement identifiées comme des anomalies.

Dans un scénario idéal de détection d'anomalies (Figure 1.2a), toute paire de points de données (x_i, x_j), où $x_i \in \mathcal{A}$ représente une observation anormale et $x_j \in \mathcal{N}$ une observation normale, devrait recevoir des scores d'anomalie nettement distincts, de sorte que $\psi(x_i) > \psi(x_j)$, sans aucun chevauchement entre leurs distributions de scores. Cette séparation permet à toute valeur située à la frontière de ces deux distributions de servir efficacement de seuil de décision (par exemple, $\tau = 0.35$).

Cependant, les scénarios réels présentent souvent des défis significatifs (Figure 1.2b). Typiquement, les distributions de scores d'anomalie pour les données normales et anormales se chevauchent, ce qui entraîne des erreurs de détection. Ce chevauchement se traduit par deux types d'erreurs : les détections ratées (anomalies mal classifiées comme normales) et les fausses alarmes (observations normales incorrectement identifiées comme anormales), comme montré dans la Figure 1.2c.

Ajuster le seuil τ permet d'affiner le processus de détection en fonction des besoins spécifiques de l'application. Bien que le seuillage ne soit pas nécessairement au centre des préoccupations de la recherche académique, il s'avère indispensable dans la pratique industrielle en tant qu'étape de post-traitement de la détection d'anomalies. Cet étalonnage peut être effectué manuellement par des experts du domaine et des analystes ou automatiquement via des mécanismes de seuillage dynamique [Ali et al., 2013; Hundman et al., 2018; Li et al., 2020a; Tayeh et al., 2022]. Cette étape est particulièrement importante pour des méthodes sensibles au seuil comme XGBOD [Zhao and Hryniewicki, 2018].

1.3 Approches

La recherche sur la détection d'anomalies a une longue histoire, qui couvre plusieurs disciplines, notamment l'ingénierie, la statistique, l'apprentissage automatique et la fouille de données. L'étude des observations aberrantes remonte à la fin du XIXe siècle, lorsque les « observations discordantes » ont été formellement définies et rigoureusement étudiées dans le domaine de la statistique [Edgeworth, 1887].

Le début du XXe siècle a vu la poursuite du développement de la détection des anomalies, notamment grâce aux contributions de Walter A. Shewhart dans les années 1930. Shewhart [1930] a développé des cartes de contrôle pour le contrôle de qualité et la maîtrise statistique des procédés, qui sont considérées comme l'un des premiers, sinon le tout premier, algorithmes de détection d'anomalies.

Au cours de la seconde moitié du XXe siècle, le domaine de la détection d'anomalies a gagné une importance croissante, en particulier dans la fouille de données. Depuis les années 1960, plusieurs techniques statistiques ont été adaptées pour identifier les « valeurs aberrantes » [Anscombe and Guttman, 1960; Box and Tiao, 1968; Grubbs, 1969; Fox, 1972; Hawkins, 1974].

La résurgence de l'intérêt pour l'apprentissage automatique dans les années 1980 a grandement stimulé le progrès de la détection d'anomalies, donnant naissance à de nombreuses nouvelles méthodes dans divers domaines tels que la sécurité informatique [Denning, 1987; Ilgun et al., 1995; Ghosh et al., 1998], la finance [Aleskerov et al., 1997; Brause et al., 1999; Donoho, 2004], et la médecine [He et al., 1997; Lin et al., 2005].

Avec l'essor de l'apprentissage profond dans les années 2010, des techniques inno-

vantes basées sur diverses architectures de réseaux de neurones ont émergé, qui varient considérablement en termes de complexité, d'exactitude et d'applicabilité [Chalapathy and Chawla, 2019; Pang et al., 2021a,b].

Des approches statistiques traditionnelles aux techniques avancées d'apprentissage profond, le domaine de la détection d'anomalies a connu une évolution remarquable au cours des dernières décennies. Dans cette section, nous explorerons ces méthodes sous plusieurs angles. Nous les classerons d'abord en fonction de leur paradigme d'apprentissage, en nous concentrant sur l'utilisation des étiquettes de vérité terrain. Ensuite, nous nous pencherons sur l'architecture des modèles, en faisant la distinction entre les approches traditionnelles d'apprentissage automatique, parfois qualifiées d'apprentissage « peu profond (*shallow*) » [Ding et al., 2019; Ruff et al., 2021; Han et al., 2022], et les méthodes plus récentes et plus complexes basées sur l'apprentissage profond. Enfin, nous examinerons en détail les techniques selon le calcul des scores d'anomalie, l'élément central de tout algorithme de détection d'anomalies. Cette analyse multidimensionnelle permettra de dresser un panorama complet des méthodes actuelles de détection d'anomalies, mettant en lumière les tendances générales de la recherche.

1.3.1 Paradigme d'apprentissage

L'apprentissage automatique, composant fondamental de l'intelligence artificielle (IA), désigne l'ensemble des méthodes et techniques qui permettent aux machines d'acquérir des connaissances à partir des données et de prendre des décisions basées sur cet apprentissage. Contrairement à la programmation traditionnelle, qui repose sur des instructions explicites pour effectuer des tâches spécifiques, l'apprentissage automatique entraîne des algorithmes pour identifier des patterns, faire des prédictions ou prendre des décisions sans être explicitement programmé pour ces tâches spécifiques [Zhou, 2021].

En ce qui concerne la détection d'anomalies, il s'agit d'entraîner des algorithmes sur des données pour créer des modèles qui capturent le comportement normal au sein d'un ensemble de données. Ces modèles identifient ensuite les déviations significatives comme des anomalies. Depuis sa renaissance dans les années 1980 après l'hiver de l'IA, l'apprentissage automatique est devenu l'approche prédominante dans le domaine de la détection d'anomalies [Chandola et al., 2009; Omar et al., 2013; Nassif et al., 2021]. Selon l'utilisation de données étiquetées pour guider la construction du modèle, nous pouvons distinguer quatre principaux paradigmes d'apprentissage :

— **Apprentissage supervisé** : Le paradigme d'apprentissage le plus élémentaire où les données d'entraînement sont entièrement étiquetées, fournissant des indications claires sur ce qui constitue la norme et ce qui relève d'une anomalie.

— **Apprentissage non supervisé** : Ce paradigme fonctionne sans données étiquetées et a traditionnellement dominé la détection d'anomalies, où les anomalies sont rarement connues ou étiquetées à l'avance.

— **Apprentissage semi-supervisé** : Ce paradigme comble le fossé entre l'apprentissage entièrement supervisé et l'apprentissage non supervisé en utilisant des données partiellement étiquetées pour guider le processus d'entraînement.

— **Apprentissage faiblement supervisé** : Ce paradigme utilise un petit nombre de données étiquetées ainsi qu'une grande quantité de données non étiquetées. L'idée est d'utiliser les données étiquetées limitées pour guider le processus

d'apprentissage tout en exploitant l'abondance de données non étiquetées pour améliorer les performances du modèle.

Il convient de noter que les termes « apprentissage semi-supervisé » et « apprentissage faiblement supervisé » n'ont pas de définitions universellement acceptées et ont des portées variablement délimitées dans le domaine de la détection d'anomalies. Différents auteurs peuvent utiliser ces termes de diverses manières. Ainsi, dans la discussion suivante, nous définirons spécifiquement ces termes et expliquerons leur portée dans le contexte de notre étude, qui peut différer de leur utilisation dans d'autres travaux sur la détection d'anomalies ou dans d'autres domaines.

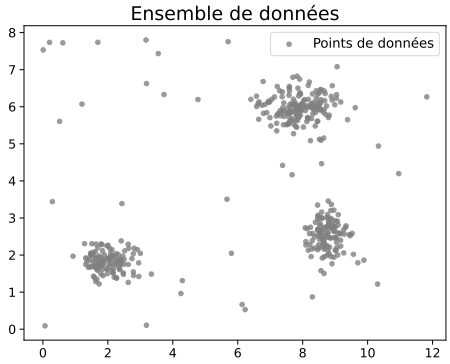
1.3.1.1 Approche supervisée

L'apprentissage supervisé est le paradigme le plus courant de l'apprentissage automatique [Jordan and Mitchell, 2015; Li, 2024], où les modèles sont entraînés à partir des données entièrement étiquetées (Figure 1.3f). Chaque échantillon d'entraînement comprend des données d'entrée ainsi que la sortie correcte correspondante, ce qui permet au modèle d'apprendre une correspondance entre les entrées et les sorties. Cette approche est typiquement utilisée pour des problèmes tels que la classification et la régression.

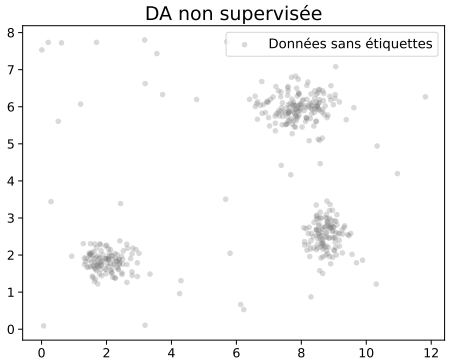
Grâce à ce haut degré de supervision, les méthodes entièrement supervisées se distinguent par leur haute exactitude de détection, notamment pour des données complexes et à haute dimension [Goernitz et al., 2013; Omar et al., 2013]. Néanmoins, lorsqu'il s'agit de détection d'anomalies, cette approche nécessite un ensemble de données volumineux et entièrement annoté, où chaque instance est exactement étiquetée comme normale ou anormale. Cette exigence pose des défis considérables, car la rareté et l'irrégularité intrinsèques des anomalies rendent difficile, voire impossible, la collecte et l'annotation d'un ensemble d'anomalies suffisamment représentatif.

Malgré ces limitations, certaines études ont tenté d'effectuer la détection d'anomalies dans un cadre entièrement supervisé. Ces efforts, cependant, ne répondent pas spécifiquement aux défis uniques de la détection d'anomalies ; en revanche, ils abordent le problème dans le cadre d'une classification binaire qui est caractérisée par un déséquilibre significatif entre le nombre d'échantillons normaux et anormaux. En général, il s'agit de construire un classificateur, soit discriminatif, soit génératif, capable de distinguer efficacement les comportements normaux des comportements déviants. Pour résoudre les problèmes liés au déséquilibre de classes et au manque d'anomalies étiquetées, les chercheurs ont mis en œuvre des stratégies telles que le développement de nouvelles fonctions de perte, le suréchantillonnage des classes minoritaires et l'utilisation de techniques d'augmentation de données [Subudhi and Panigrahi, 2015; Hassan and Abraham, 2016; Wang et al., 2017; Yamanaka et al., 2019; Liu et al., 2021a]. En dépit de ces efforts, l'adoption de l'apprentissage entièrement supervisé dans la détection d'anomalies reste très limitée en raison de problèmes tels que les anomalies hors distribution (dérives conceptuelles et nouveautés), le coût élevé de l'annotation et les risques de surapprentissage [Chandola et al., 2009; Aggarwal, 2017a; Han et al., 2022].

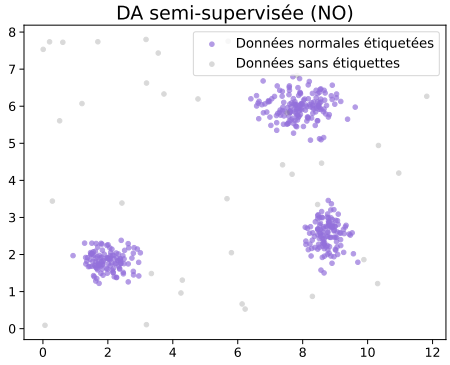
L'apprentissage entièrement supervisé s'avère très efficace dans la détection d'anomalies [Goernitz et al., 2013; Aggarwal, 2017a]. Par conséquent, il reste une option pertinente pour des contextes disposant de données étiquetées en quantité suffisante, notamment pour des anomalies bien comprises ou prévisibles. De plus,



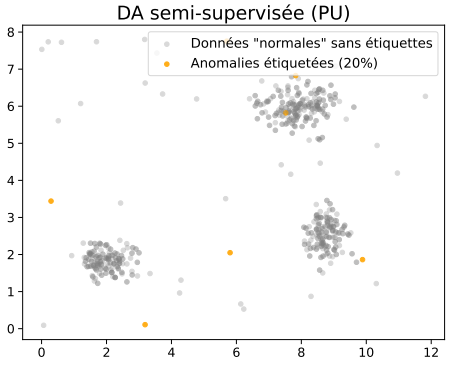
(a) Ensemble de données



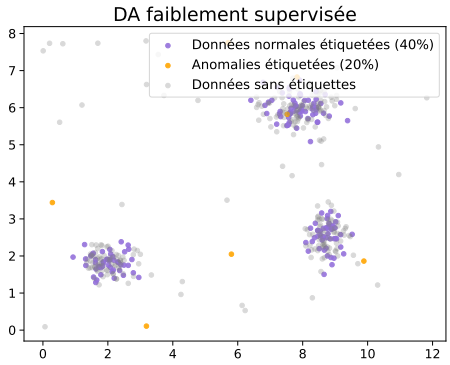
(b) DA non supervisée



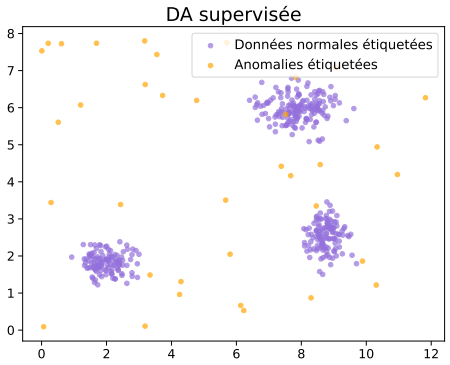
(c) DA semi-supervisée (*Normal-Only*)



(d) DA semi-supervisée (*Positive-Unlabeled*)



(e) DA faiblement supervisée



(f) DA supervisée

FIGURE 1.3 – Paradigmes d'apprentissage de la détection d'anomalies (DA).

dans des scénarios où les données étiquetées sont rares ou difficiles à obtenir, une des grandes tendances de la recherche récente est d'introduire un certain niveau de supervision, comme dans les approches semi-supervisées et faiblement supervisées, pour améliorer les modèles de détection d'anomalies [Goernitz et al., 2013; Zhao and Hryniewicki, 2018; Ruff et al., 2020; Pang et al., 2020; Villa-Pérez et al., 2021; Han et al., 2022; Jiang et al., 2023]. Ces approches visent à tirer parti des données étiquetées limitées tout en bénéficiant de la flexibilité des méthodes non supervisées. Tou-

tefois, comme l'indiquent les travaux existants [Goernitz et al., 2013; Kawachi et al., 2018], pour détecter efficacement les anomalies inconnues tout comme connues, ces méthodes, au lieu de s'appuyer uniquement sur le cadre de l'apprentissage entièrement supervisé, devraient se fonder sur le paradigme de l'apprentissage non supervisé, qui est privilégié dans la détection d'anomalies.

1.3.1.2 Approche non supervisée

Contrairement aux méthodes supervisées qui s'appuient sur des données étiquetées pour apprendre à distinguer les instances anormales des instances normales, la détection d'anomalies non supervisée établit la norme exclusivement à partir de données non étiquetées (Figure 1.3b). Ce paradigme repose sur l'hypothèse générale que les instances normales sont beaucoup plus nombreuses que les anomalies au sein de l'ensemble de données. Si cette hypothèse s'avère fausse, autrement dit le taux d'anomalies est sous-estimé, ces techniques peuvent souffrir d'un taux de fausse alarme élevé [Chandola et al., 2009].

Les méthodes non supervisées ne nécessitent pas de données d'entraînement étiquetées. Leur objectif principal est plutôt de découvrir des patterns cachés et des structures intrinsèques au sein des données. Contrairement à l'apprentissage supervisé, qui vise à développer des modèles capables de généraliser sur de nouvelles instances non vues auparavant, la détection d'anomalies non supervisée utilise généralement une approche d'inférence transductive [Ding et al., 2021]. Cette approche se concentre sur la prédiction directe sur les données d'entrée spécifiques, ce qui s'aligne bien avec la nature imprévisible et hautement contextuelle des anomalies. Il est également possible d'adapter ces méthodes à l'inférence inductive en sérialisant simplement les paramètres appris pendant l'entraînement pour les appliquer à de nouvelles données [Han et al., 2022].

Étant donné la nature dynamique et irrégulière des anomalies, l'apprentissage non supervisé est particulièrement adapté à la détection d'anomalies. Ne dépendant pas de connaissances préalables sur ce qui constitue une anomalie, les méthodes non supervisées se révèlent idéales pour identifier des patterns inhabituels au fur et à mesure de leur apparition. Par conséquent, la détection d'anomalies est traditionnellement considérée comme une tâche typiquement non supervisée et a été principalement explorée dans ce cadre [Chandola et al., 2009; Pimentel et al., 2014; Aggarwal, 2017a; Pang et al., 2021b].

Dans la littérature, de nombreuses méthodes de détection d'anomalies non supervisées ont été développées en se basant sur différentes hypothèses de distribution des données et théories sous-jacentes, conduisant à diverses approches pour calculer les scores d'anomalie. Ces méthodes englobent un large éventail de techniques, notamment les méthodes basées sur les statistiques, sur la proximité et sur la reconstruction, entre autres.

Par exemple, les méthodes à base de proximité supposent que les instances normales sont situées dans des régions denses de l'espace de données, tandis que les anomalies sont isolées et se trouvent dans des régions peu peuplées. Une méthode typique basée sur la proximité est l'approche des k plus proches voisins (*K-Nearest Neighbors*, KNN) [Zhao et al., 2018; Gu et al., 2019], où une instance est considérée comme anormale si elle est éloignée de ses voisins les plus proches dans l'espace des caractéristiques.

En effet, la plupart des méthodes semi-supervisées et faiblement supervisées peuvent être adaptées au paradigme non supervisé en utilisant une partie des données non étiquetées pour l'apprentissage [Akcay et al., 2019; Pang et al., 2020; Manolache et al., 2021; Pang et al., 2023]. Cette adaptation repose sur l'hypothèse que les données de test comporteront très peu d'anomalies et que le modèle développé pendant la phase d'entraînement restera efficace malgré la présence de ces anomalies peu nombreuses [Chandola et al., 2009].

1.3.1.3 Approche semi-supervisée

Historiquement, la détection d'anomalies a été dominée par des méthodes non supervisées [Markou and Singh, 2003a,b; Chandola et al., 2009; Goldstein and Uchida, 2016; Aggarwal, 2017a; Pang et al., 2020]. Cependant, l'introduction d'un certain degré de supervision a démontré une amélioration significative des performances, marquant une tendance notable dans les recherches récentes [Zhao and Hryniewicki, 2018; Kiran et al., 2018; Villa-Pérez et al., 2021; Han et al., 2022; Jiang et al., 2023]. L'apprentissage semi-supervisé représente une avancée significative à cet égard en intégrant un niveau modéré de supervision. Cette méthode trouve un équilibre entre l'apprentissage entièrement supervisé et non supervisé, en utilisant des données partiellement étiquetées dans le processus d'entraînement.

L'apprentissage semi-supervisé est une approche hybride qui exploite à la fois des données étiquetées et non étiquetées. Théoriquement, il peut s'agir de toute combinaison possible de données normales et anormales, mais dans le contexte de la détection d'anomalies, le paradigme semi-supervisé consiste généralement à guider le processus d'entraînement avec des classes partiellement observées. Typiquement, les données contiennent une classe normale et plusieurs classes anormales. Le paradigme semi-supervisé peut se concentrer sur l'un ou l'autre scénario en fonction des données étiquetées disponibles : échantillons de comportements normaux ou instances d'anomalies.

Paradigme d'apprentissage *Normal-Only* (NO) La forme la plus courante de détection d'anomalies semi-supervisée est le paradigme *Normal-Only* (NO) [Chandola et al., 2009; Aggarwal, 2017a; Akcay et al., 2019], où seuls les exemples normaux sont étiquetés et aucune instance anormale n'est observée pendant la phase d'apprentissage (Figure 1.3c). Ce scénario s'aligne étroitement avec les techniques de classification à classe unique (*One-Class Classification*, OCC) [Tax, 2001; Khan and Madden, 2014; Pimentel et al., 2014; Perera et al., 2021; Seliya et al., 2021] dans l'apprentissage automatique, où l'objectif est de modéliser le comportement normal exclusivement à partir des données d'entraînement normales, puis d'utiliser le modèle pour identifier (et rejeter) les anomalies potentielles lors de l'inférence.

Les méthodes traditionnelles dans cette catégorie incluent des techniques basées sur la région, telles que les machines à vecteurs de support à classe unique (*One-Class Support Vector Machines*, OCSVM) [Schölkopf et al., 1999] et la description des données par vecteurs de support (*Support Vector Data Description*, SVDD) [Tax, 2001], ainsi que leurs variantes [Duong et al., 2015; Ruff et al., 2018; Wang and Cherian, 2019; Liu and Gryllias, 2020; Yang et al., 2021; Mukherjee et al., 2022], qui se concentrent sur la définition d'une frontière autour des données normales. Les méthodes non supervisées basées sur la proximité, comme les KNN [Agarwal and Sureka, 2015; Daneshpazhouh and Sami, 2015; Dang et al., 2015; Song et al.,

2017] et le *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN) [Zhou et al., 2017; Castro Gertrudes et al., 2019; Deng and Brown, 2022], peuvent également être adaptées pour exploiter des données partiellement étiquetées.

Pour les modèles profonds, les approches basées sur les réseaux antagonistes génératifs (*Generative Adversarial Networks*, GANs) comme AnoGAN [Schlegl et al., 2017, 2019; Yap, 2020] et Ganomaly [Akcay et al., 2019; Luo et al., 2021] fonctionnent généralement en mode semi-supervisé, visant à générer des distributions de données normales et à identifier les déviations comme des anomalies. De plus, des réseaux de neurones pré-entraînés comme *One-Class Convolutional Neural Network* (OCCNN) [Oza and Patel, 2019] et *Multi-layer One-Class ClassificAtion* (MOCCA) [Massoli et al., 2022] ont été proposés pour détecter des anomalies dans le cadre d'OCC via l'apprentissage par transfert. Récemment, l'apprentissage *few-shot* a été introduit dans la détection d'anomalies semi-supervisée, visant à modéliser efficacement le comportement normal avec un nombre minimal d'exemples normaux [Frikha et al., 2021; Belton et al., 2023].

Ces méthodes bénéficient d'un niveau de supervision plus élevé que les techniques non supervisées, aboutissant souvent à une exactitude de prédiction améliorée. Elles offrent également un champ d'application plus large par rapport aux techniques supervisées, car elles ne nécessitent pas d'étiquettes pour les classes d'anomalies. Il convient toutefois de noter que la frontière entre les méthodes non supervisées et les méthodes NO semi-supervisées est souvent floue, étant donné l'hypothèse de base de la détection d'anomalies non supervisée. En pratique, de nombreuses techniques non supervisées peuvent être adaptées ou appliquées directement en mode d'inférence inductive pour être utilisées dans un cadre semi-supervisé [Chandola et al., 2009; Aggarwal, 2017a].

Paradigme d'apprentissage *Positive-Unlabeled* (PU) Dans certains scénarios, seules quelques instances des classes anormales sont disponibles. Si la distribution des données satisfait à l'hypothèse que les données normales sont nettement prédominantes, ces scénarios sont souvent abordés dans le cadre du paradigme d'apprentissage *Positive-Unlabeled* (PU) [Aggarwal, 2017a; Bekker and Davis, 2020; Jaskie and Spanias, 2022]. L'apprentissage PU utilise un nombre limité d'instances des classes positives (anormales) ainsi qu'une grande quantité de données non étiquetées, en supposant que les données non étiquetées sont essentiellement constituées d'instances normales, avec une très faible proportion d'anomalies (contamination).

Il convient de préciser ici deux points. Premièrement, dans notre étude, nous considérons les anomalies comme la classe positive à partir d'une perspective traditionnelle, ce qui est en ligne avec la pratique courante en détection d'anomalies [Chandola et al., 2009; Pimentel et al., 2014]. Cependant, certaines recherches utilisent l'apprentissage PU inversement, où la classe positive désigne un grand nombre de données normales étiquetées [Zhang et al., 2017, 2018, 2019; Mu et al., 2021a]. Dans de tels cas, ces méthodes montrent peu de différence par rapport aux approches semi-supervisé NO en pratique [Aggarwal, 2017a]. Deuxièmement, dans la littérature plus récente sur la détection d'anomalies, l'apprentissage PU est parfois discuté sous l'apprentissage faiblement supervisé en raison de son utilisation des étiquettes incomplètes [Ortega Vázquez et al., 2023; Wittkopp et al., 2023]. Dans cette étude, pour établir une délimitation claire entre l'apprentissage semi-supervisé et l'apprentissage faiblement supervisé, nous adoptons une définition plus stricte de la supervi-

sion faible, classant ainsi l'apprentissage PU comme semi-supervisé avec des classes partiellement observées.

Portée de la détection d'anomalies semi-supervisée Dans notre analyse, les méthodes de détection d'anomalies semi-supervisées sont principalement catégorisées en approches *Normal-Only* et *Positive-Unlabeled*. Les méthodes qui utilisent un nombre limité d'échantillons étiquetés pour les deux classes (normales et anormales), ainsi qu'un ensemble plus large de données non étiquetées, sont examinées dans le cadre de l'apprentissage faiblement supervisé pour maintenir une distinction claire entre les paradigmes.

1.3.1.4 Approche faiblement supervisée

La détection d'anomalies faiblement supervisée constitue le dernier développement de la tendance à incorporer différents niveaux de supervision dans les techniques de détection d'anomalies. Ce paradigme est particulièrement efficace dans les scénarios où il existe une quantité limitée de données étiquetées, combinée à un grand volume de données non étiquetées. En général, il existe trois types de supervision faible : la supervision incomplète, imprécise et incorrecte [Zhou, 2018].

— **Supervision incomplète** : Il s'agit de la forme de supervision faible la plus courante dans la détection d'anomalies, où seul une proportion limitée des données d'entraînement est étiquetée. Le coût élevé de l'annotation des données limite souvent la disponibilité des ensembles de données entièrement étiquetés, ce qui rend la supervision incomplète particulièrement pertinente pour la détection d'anomalies.
— **Supervision imprécise** : Dans ce cas, les étiquettes sont fournies uniquement à gros grains. Bien que courante dans d'autres domaines, l'annotation imprécise n'est généralement pas un problème majeur dans la détection d'anomalies.
— **Supervision incorrecte** : Il s'agit des scénarios dans lesquels certaines des étiquettes fournies ne sont pas tout à fait correctes. Les applications de détection d'anomalies abordent souvent ce problème en développant des algorithmes résistants au bruit des étiquettes, plutôt qu'en intégrant une supervision faible.

En raison de sa prédominance et de sa forte pertinence, la plupart des travaux sur la détection d'anomalies faiblement supervisée se concentrent sur la supervision incomplète (Figure 1.3e). L'objectif central est d'améliorer les capacités de détection d'anomalies en utilisant efficacement des données étiquetées limitées, en particulier les anomalies.

L'une des principales approches faiblement supervisées est l'apprentissage de la représentation des caractéristiques des anomalies. Cette approche vise à améliorer l'apprentissage de la représentation en utilisant des informations partiellement disponibles. L'idée principale est d'apprendre des représentations de données enrichies et spécifiques aux anomalies simultanément à partir de données normales et anormales, permettant ainsi de créer un profil plus fidèle et complet de la norme au sein des données.

Par exemple, les méthodes comme *Extreme Boosting Based Outlier Detection* (XGBOD) [Zhao and Hryniewicki, 2018] utilisent des algorithmes de détection d'anomalies non supervisés comme extracteurs de caractéristiques. Elles enrichissent les ca-

ractéristiques originales avec des scores d'anomalie prédits par ces algorithmes, en utilisant à la fois des données étiquetées et non étiquetées. Cela permet une représentation plus exacte du comportement normal, améliorant ainsi les capacités de détection du système. De même, *Deep Semi-Supervised Anomaly Detection* (DeepSAD) [Ruff et al., 2020] se construit sur la méthode non supervisée profonde *Deep Support Vector Data Description* (DeepSVDD) [Ruff et al., 2018] en introduisant un nombre limité d'instances étiquetées lors de l'apprentissage. Cette méthode impose une pénalité sur l'inverse de la distance des caractéristiques d'anomalie dans l'espace de plongements, garantissant que les anomalies sont mappées plus loin du centre d'une hypersphère de plongements. Une autre méthode, REPEN [Pang et al., 2018], utilise un petit nombre d'anomalies étiquetées et applique un échantillonnage de triplets pour apprendre des représentations de données expressives. Ce processus implique la création de triplets de haute qualité qui aident le modèle à distinguer les données normales des anomalies, améliorant ainsi la qualité des représentations apprises.

Des efforts plus récents ont exploré l'utilisation de techniques telles que les GANs et l'apprentissage *few-shot* pour améliorer l'entraînement du modèle par l'augmentation des données dans le cadre faiblement supervisé. Par exemple, *Dual Multiple Generative Adversarial Networks* (Dual-MGAN) [Li et al., 2022b] intègre plusieurs GANs pour construire des distributions de référence et augmenter les données, renforçant ainsi les performances de détection d'anomalies. De même, Kale and Thing [2023] proposent un cadre d'apprentissage *few-shot* qui enrichit les données d'entraînement en générant des échantillons augmentés par des combinaisons de triplets d'anomalies étiquetées et d'échantillons non étiquetés, suivies d'un apprentissage de représentation et d'une régression ordinale pour améliorer la détection d'anomalies.

Ces méthodes contribuent collectivement à atténuer les difficultés causées par la connaissance incomplète des anomalies en maximisant l'utilité des données étiquetées disponibles. En améliorant la représentation des données et en incorporant l'augmentation des données, la détection d'anomalies faiblement supervisée offre une approche robuste pour identifier les anomalies, même lorsque l'étiquetage complet est irréalisable.

1.3.2 Architecture de modèle

La détection d'anomalies emploie diverses architectures de modèles pour distinguer efficacement les anomalies des données normales. Un aspect important dans l'analyse de l'architecture des modèles est la distinction entre les modèles « peu profond (*shallow*) » et « profonds (*deep*) » [Ding et al., 2019; Ruff et al., 2021; Han et al., 2022], en fonction de leur complexité et l'utilisation de réseaux de neurones. Cette distinction est essentielle car elle influence la performance, l'applicabilité et les exigences opérationnelles du modèle. Comprendre les différences entre ces modèles permet de choisir l'approche appropriée en fonction des exigences spécifiques et des contraintes des applications ciblées.

1.3.2.1 Modèles peu profonds

Dans le contexte de la détection d'anomalies, les modèles peu profonds désignent généralement les algorithmes d'apprentissage automatique qui n'utilisent pas de réseaux de neurones ou qui emploient des réseaux de neurones peu profonds avec une seule couche cachée [Markou and Singh, 2003b]. Ces modèles sont souvent plus

simples et linéaires, sans multiples couches de traitement ou d'abstraction. Ils se caractérisent par leur transparence, leur facilité de mise en œuvre et leurs faibles exigences en termes de ressources informatiques par rapport aux modèles profonds. Dans cette thèse, ces modèles sont également désignés comme des modèles d'apprentissage automatique traditionnels.

Les modèles peu profonds englobent une variété de méthodes. La plupart des méthodes non supervisées basées sur la proximité et les statistiques entrent dans cette catégorie. Les modèles à base de proximité, tels que le *Local Outlier Factor* (LOF) [Breunig et al., 2000] et le K-means [Hartigan and Wong, 1979; Munz et al., 2007], ainsi que leurs variantes, détectent les anomalies en s'appuyant sur la densité ou le clustering des points de données [He et al., 2003; Papadimitriou et al., 2003; Li et al., 2011; Chauhan and Shukla, 2015; Pevný, 2016; Wazid and Das, 2016; Na et al., 2018; Wang et al., 2020; Boukela et al., 2021; Ripan et al., 2021; Yong et al., 2022]. Les techniques telles que les modèles de mélange gaussien (*Gaussian Mixture Models*, GMM) [Markou and Singh, 2003a; Miljković, 2010; Pimentel et al., 2014] et le modèle *Histogram-based Outlier Score* (HBOS) [Goldstein and Dengel, 2012] identifient les anomalies en modélisant la distribution statistique des données.

Certains modèles simples à base de région et des premières techniques à base de reconstruction appartiennent également à ce groupe. Les modèles de classification à classe unique (OCC) comme OCSVM et SVDD créent une frontière autour de la région des données normales pour identifier les anomalies. Les méthodes de réduction de dimension telles que l'analyse en composantes principales (*Principal Component Analysis*, PCA) [Shyu et al., 2003; Brauckhoff et al., 2009; Hoang and Nguyen, 2018] et ses variantes [Hoffmann, 2007; Pascoal et al., 2012; Siwach and Mann, 2022] compriment les données , puis les reconstruisent pour trouver des anomalies en mesurant les erreurs de reconstruction.

Les développements récents des modèles peu profonds visent à relever des défis spécifiques. Par exemple, les méthodes basées sur les statistiques comme le modèle *Copula-Based Outlier Detection* (COPOD) [Li et al., 2020b] et le modèle *Empirical Cumulative Distribution Functions for Outlier Detection* (ECOD) [Li et al., 2022c] sont optimisées pour traiter la malédiction de la dimensionnalité. Pour améliorer les performances, les modèles à base l'apprentissage ensembliste, comme XGBOD, combinent plusieurs modèles peu profonds pour améliorer la robustesse et l'exactitude de la détection.

Les modèles peu profonds sont connus pour leurs exigences moindres en termes de ressources informatiques, ce qui les rend idéaux pour les applications en temps réel et les scénarios avec des ressources limitées. Un autre avantage majeur des modèles peu profonds est leur haute interprétabilité, ce qui les rend particulièrement intéressants pour les contextes où il est important de maîtriser le processus de prise de décision du modèle. Cependant, leur efficacité diminue avec l'augmentation de la dimensionnalité des données, et ils ne s'adaptent souvent pas bien aux ensembles de données plus volumineux.

1.3.2.2 Modèles profonds

Les modèles profonds dans la détection d'anomalies sont définis par l'utilisation de réseaux de neurones profonds avec plusieurs couches cachées. Ces modèles sont capables d'exécuter des transformations complexes et non linéaires, leur permettant

d'apprendre des patterns et des représentations complexes à partir des données. Ils sont donc particulièrement efficaces pour traiter des tâches de détection d'anomalies complexes et de haute dimension.

Les approches profondes les plus courantes sont les méthodes d'OCC profondes et les modèles basés sur des variantes d'auto-encodeurs et des GANs. Les méthodes d'OCC profondes utilisent souvent des techniques basées sur les régions, notamment des modèles tels que DeepSVDD, DeepSAD et One-Class Neural Network (OC-NN) [Chalapathy et al., 2019]. Comme les modèles peu profonds à base de région, ils créent des frontières complexes autour des régions de données normales dans des espaces de haute dimension. Les méthodes basées sur des autoencodeurs, à l'aide de leur structure codeur-décodeur, apprennent à compresser et reconstruire les données, ce qui permet d'identifier les anomalies en utilisant les erreurs de reconstruction. Les méthodes basées sur les GANs, telles que AnoGAN, Ganomaly et *Adversarially Learned Anomaly Detection* (ALAD) [Zenati et al., 2018], utilisent les capacités génératives des GANs pour modéliser la distribution des données normales et détecter les anomalies comme des écarts par rapport à cette distribution.

Les modèles ensemblistes bénéficient également des réseaux de neurones profonds. Ces modèles intègrent des réseaux de neurones dans des modèles ensemblistes peu profonds pour améliorer la représentation des caractéristiques ou combinent plusieurs modèles de base profonds pour obtenir une performance supérieure [Tang et al., 2020; Tsogbaatar et al., 2021; Sarvari et al., 2021; Khan and Haroon, 2022].

Les développements plus récents impliquent l'intégration de réseaux de neurones dans des cadres semi-supervisés et faiblement supervisés pour maximiser l'utilité des données étiquetées limitées. Les méthodes à base de l'apprentissage des scores, telles que DevNet [Pang et al., 2019] et PreNet [Pang et al., 2023], illustrent cette tendance. DevNet emploie une approche d'apprentissage de bout en bout en entraînant directement le réseau à produire des scores d'anomalie, évitant ainsi les inconvénients causés par le processus de détection traditionnel en deux étapes. PreNet, en revanche, utilise des réseaux de neurones pour l'apprentissage des scores par paires, ce qui permet d'affiner sa capacité à distinguer efficacement les instances normales et anormales. En termes d'apprentissage de représentation, des méthodes comme *Feature Encoding with AutoEncoders for Weakly Supervised Anomaly Detection* (FEAWAD) [Zhou et al., 2022b] utilisent des auto-encodeurs pour encoder les données dans un espace de plus basse dimension qui discrimine mieux entre les patterns normaux et anormaux, améliorant ainsi les capacités de détection en utilisant efficacement les données étiquetées et non étiquetées.

Les modèles profonds excellent dans le traitement des données à haute dimension et dans l'apprentissage de petterns complexes que les modèles peu profonds pourraient rater. Ils sont particulièrement efficaces dans les applications impliquant des images, de l'audio ou des données séquentielles [Nassif et al., 2021; Ruff et al., 2021]. Cependant, les modèles profonds présentent des inconvénients majeurs, notamment des exigences informatiques élevées et des défis en termes d'interprétabilité. L'entraînement et le déploiement des modèles profonds nécessitent une puissance de calcul et une mémoire substantielles, ce qui peut être prohibitif dans les environnements avec des ressources limitées. De plus, leur complexité entraîne souvent une transparence réduite, posant des problèmes critiques dans les domaines où l'explicabilité est essentielle.

En conclusion, les modèles profonds surpassent les modèles traditionnels peu profonds dans de nombreux cas, ce qui entraîne une tendance générale à favoriser l'introduction de réseaux de neurones profonds lorsque les conditions le permettent [Chalapathy and Chawla, 2019; Al-amri et al., 2021; Nassif et al., 2021; Ruff et al., 2021; Han et al., 2022; Landauer et al., 2023]. Cependant, les modèles peu profonds et profonds présentent tous deux des avantages et des inconvénients distincts, ce qui rend le choix entre eux dépendant des besoins spécifiques de l'application, de la complexité des données et des ressources informatiques disponibles. Comprendre ces différences est crucial pour concevoir des systèmes de détection d'anomalies efficaces, adaptés aux défis uniques présentés par différents ensembles de données et environnements opérationnels.

1.3.3 Scores d'anomalie

Comme discuté dans la section §1.2, la plupart des techniques de détection d'anomalies en fouille de données se concentrent sur le calcul des scores d'anomalie. Ces scores indiquent de manière quantitative le degré de déviation par rapport au comportement normal. Typiquement, le calcul des scores suit un processus en deux étapes : d'abord l'établissement d'une norme, puis l'évaluation de la déviation par rapport à cette norme. Selon la technique utilisée, ce processus peut être fondé sur différentes hypothèses et guidé par divers cadres théoriques. Par conséquent, ces techniques peuvent être catégorisées en plusieurs groupes, y compris les cadres traditionnels tels que les méthodes basées sur les statistiques, la proximité, la région, la reconstruction et la théorie de l'information, ainsi que les tendances plus récentes comme les méthodes ensembliste et l'apprentissage des scores de bout en bout [Markou and Singh, 2003a,b; Patcha and Park, 2007; Chandola et al., 2009; Miljković, 2010; Pimentel et al., 2014; Agrawal and Agrawal, 2015; Aggarwal, 2017a; Pang et al., 2021b; Ruff et al., 2021; Samariya and Thakkar, 2023]. Cette section explorera les méthodes de détection d'anomalies du point de vue du calcul des scores d'anomalie.

1.3.3.1 Approche basée sur les statistiques

Les méthodes basées sur les statistiques sont les premières approches pour la détection d'anomalies et ont dominé le domaine avant l'avènement des techniques d'apprentissage automatique. Ces méthodes utilisent des modèles probabilistes et statistiques pour identifier les anomalies en supposant que les instances de données normales se trouvent dans les régions à haute probabilité d'un modèle stochastique, tandis que les anomalies se trouvent dans les régions de probabilité plus faible [Markou and Singh, 2003a; Clifton et al., 2009; Chandola et al., 2009; Aggarwal, 2017a; Samariya and Thakkar, 2023]. Pour y parvenir, un modèle statistique est d'abord ajusté aux données afin de représenter le comportement normal. Les instances ayant une faible probabilité d'être générées par ce modèle sont ensuite signalées comme anomalies par des tests d'inférence statistique. Les méthodes à base de statistiques peuvent être paramétriques ou non paramétriques.

Les méthodes paramétriques supposent que les données suivent une distribution spécifique, typiquement gaussienne, et utilisent des paramètres (par exemple, la moyenne et la variance) dérivés des données pour définir cette distribution. Un exemple classique est l'utilisation du test de Grubbs dans les modèles gaussiens [Grubbs, 1969; Solak, 2009; Jain, 2010; Urvoy and Autrusseau, 2014; Adikaram et al.,

2015], où les données sont supposées suivre une distribution gaussienne. La moyenne et la variance des données sont estimées à l'aide d'estimation du maximum de vraisemblance (*Maximum Likelihood Estimation*, MLE), et la distance d'un point de données par rapport à la moyenne, mesurée en écarts-types, sert de score d'anomalie. Le test de Grubbs identifie les anomalies en évaluant le degré de déviation d'un point de données par rapport à la moyenne par rapport à l'écart-type. Une autre approche paramétrique concerne des modèles de distribution de mélange, qui utilisent un mélange de plusieurs distributions paramétriques pour modéliser les données. Une méthode courante dans cette catégorie est le modèle de mélange gaussien (*Gaussian Mixture Model*, GMM) [Hollier and Austin, 2002; Bahrololum and Khaleghi, 2008; Laxhammar et al., 2009; Attar et al., 2014; Qu et al., 2021], qui suppose que les données sont générées à partir d'un mélange de distributions gaussiennes. Dans les GMM, les données sont approximées par une combinaison de plusieurs distributions gaussiennes, avec des paramètres affinés par l'algorithme espérance-maximisation (*Expectation-Maximization*, EM). Chaque point de données reçoit une probabilité d'appartenance à chaque composant gaussien, et les points ayant de faibles probabilités à travers tous les composants sont déclarés comme anomalies.

En revanche, les méthodes non paramétriques ne reposent pas sur des hypothèses concernant la distribution des données mais utilisent la structure inhérente des données pour détecter des anomalies. Une technique non paramétrique de base mais puissante est les histogrammes [Kind et al., 2009; Goldstein and Dengel, 2012; Xie et al., 2012; Wang et al., 2019b; Derhab et al., 2022], où les données sont segmentées en classes, et les fréquences d'occurrence dans chaque classe sont comptées. Les anomalies sont généralement les points qui apparaissent dans des classes avec des fréquences exceptionnellement basses, suggérant une irrégularité par rapport à la norme. Une méthode courante basée sur les histogrammes est l'*Histogram-based Outlier Score* (HBOS) [Goldstein and Dengel, 2012; Paulauskas and Baskys, 2019; Wang et al., 2019b], qui simplifie le calcul en évaluant la densité de chaque classe par un histogramme univarié. Cela permet une détection rapide et efficace des anomalies, même dans de grands ensembles de données.

Une autre technique, l'estimation de la densité par noyau (*Kernel Density Estimation*, KDE) [Latecki et al., 2007; Laxhammar et al., 2009; Gao et al., 2011; Schubert et al., 2014; Tang and He, 2017; Zheng et al., 2017; Liu et al., 2020a; Shylendra et al., 2020], estime la fonction de densité de probabilité des données en utilisant un noyau, souvent gaussien, sans supposer de modèle paramétrique spécifique. La KDE applique ce noyau à chaque point de données, avant d'agréger les résultats pour former une estimation lisse de la densité. Les points de données dans des régions de faible densité estimée sont déclarés comme anomalies, ce qui offre une méthode flexible et robuste pour identifier les anomalies dans des ensembles de données divers.

L'un des développements récents se concentre sur la résolution de la « malédiction de la dimensionnalité (*curse of dimensionality*) » , qui reste un défi majeur pour la détection d'anomalies à base de statistiques [Thudumu et al., 2020; Samariya and Thakkar, 2023]. Le modèle *Copula-Based Outlier Detection* (COPOD) [Li et al., 2020b] exploite les copules pour modéliser les dépendances entre les variables, ce qui le rend efficace dans des environnements à haute dimensionnalité. La méthode *Empirical-Cumulative-distribution-based Outlier Detection* (ECOD) [Li et al., 2022c] calcule la fonction de distribution cumulative empirique pour chaque dimension séparément, puis les agrège pour estimer les probabilités de queue conjointes, ce qui

permet d'identifier efficacement les anomalies dans les données à haute dimension en se concentrant sur les queues de la distribution à travers plusieurs dimensions.

En outre, les approches hybrides sont de plus en plus utilisées. Elles améliorent les capacités de détection en combinant des méthodes statistiques avec d'autres techniques d'apprentissage automatique, telles que l'intégration des GMM avec des méthodes de clustering ou basées sur la densité, des techniques d'OCC, et des modèles profonds comme les auto-encodeurs [Li et al., 2016; Tang and He, 2017; Zong et al., 2018; Liu et al., 2019a, 2020a; An et al., 2022; Huang et al., 2022; Lang et al., 2022; Wang et al., 2022; Ait-Saada and Nadif, 2023]. Avec la montée de l'apprentissage ensembliste dans la détection d'anomalies, les modèles basés sur les statistiques sont également devenus des composants essentiels des méthodes ensemblistes [Wang et al., 2019b; An et al., 2022; Singh and Kane, 2022].

La principale force des méthodes statistiques réside dans leur base théorique solide, qui fournit une interprétation probabiliste claire des anomalies et est très efficace dans les scénarios où la distribution des données peut être présumée ou bien approximée. Ces méthodes sont efficaces dans de nombreuses applications réelles grâce à leurs fondements mathématiques exacts. Cependant, leur efficacité dépend de l'exactitude de la distribution supposée ; des erreurs d'estimation de la distribution sous-jacente peuvent nuire considérablement la performance de la détection. De plus, les méthodes paramétriques ont souvent des difficultés à traiter des données complexes et à haute dimension, où il est difficile d'estimer précisément la distribution réelle.

1.3.3.2 Approche basée sur la proximité

Les méthodes basées sur la proximité sont un groupe de techniques simples largement utilisées dans le domaine de la détection d'anomalies. Reposant sur des concepts de distance, de densité ou de clustering pour définir et détecter les anomalies, ces méthodes sont particulièrement appréciées pour leur simplicité, leur interprétabilité et leur efficacité [Chandola et al., 2009; Duan et al., 2009; Pimentel et al., 2014; Tran et al., 2016; Aggarwal, 2017a,b]. L'idée centrale de ces méthodes est que les points de données normaux se trouvent généralement dans des voisinages densément peuplés, tandis que les anomalies se caractérisent par une proximité faiblement peuplée, qui indique une distance importante par rapport aux autres points de données.

Une des techniques les plus courantes dans la détection d'anomalies à base de proximité est la méthode des plus proches voisins. Cette approche calcule la distance de chaque point de données à ses plus proches voisins et utilise cette distance pour déterminer si un point de données constitue une anomalie. Par exemple, la méthode des k plus proches voisins (*K-Nearest Neighbors*, KNN) [Ramaswamy et al., 2000; Chen et al., 2010; Dang et al., 2015; Lei et al., 2020; Ying et al., 2021; Zhou et al., 2022a] calcule la distance de chaque point de données à son $k^{\text{ième}}$ plus proche voisin et utilise cette distance comme score d'anomalie. Les points de données ayant des distances de k plus proches voisins plus grandes sont considérés comme des anomalies potentielles. Cette méthode est non seulement couramment utilisée en soi, mais elle sert également d'élément de base pour des algorithmes et modèles plus sophistiqués [Xie et al., 2013; Agarwal and Sureka, 2015; Song et al., 2017; Zhao and Hryniewicki, 2018; Gu et al., 2019; Sarmadi and Karamodin, 2020; Chen et al., 2021b; Singh and Kane, 2022].

Les méthodes basées sur le clustering fonctionnent sur l'hypothèse que les instances de données normales appartiennent à un certain cluster dans les données, tandis que les anomalies n'appartiennent à aucun cluster. En général, ces méthodes appliquent d'abord des algorithmes de clustering tels que K-means [Hartigan and Wong, 1979; Likas et al., 2003; Munz et al., 2007; Wazid and Das, 2016; Ripan et al., 2021], *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN) [Ester et al., 1996; Duan et al., 2009; Çelik et al., 2011; Thang and Kim, 2011; Sheridan et al., 2020; Jain et al., 2022] et *Ordering Points To Identify the Clustering Structure* (OPTICS) [Ankerst et al., 1999; Breunig et al., 1999; Wang et al., 2019c; Zeufack et al., 2021; Subudhi and Panigrahi, 2022] pour identifier les régions ou clusters denses au sein des données. Ensuite, elles calculent un score d'anomalie en fonction du degré d'intégration de chaque point de données dans ces clusters. Les points qui ne s'intègrent pas bien sont signalés comme des anomalies. Cette approche est particulièrement utile dans les ensembles de données où des groupements naturels de points de données peuvent être identifiés.

En revanche, les méthodes basées sur la densité ne partitionnent pas les points de données mais plutôt l'espace des données [Aggarwal, 2017a]. Elles définissent la densité locale d'un point de données comme le nombre d'autres points dans une région spécifiée autour de ce point de données. Les valeurs de densité locale sont ensuite converties en scores d'anomalie. Par exemple, *Local Outlier Factor* (LOF) [Breunig et al., 2000; Tang et al., 2002; Huang et al., 2013; Barrett et al., 2019; Song and Suh, 2019; Boukela et al., 2021; Rousseau et al., 2021], la méthode la plus couramment utilisée dans cette catégorie, attribue un score d'anomalie à chaque point de données en fonction de sa densité locale par rapport à celle de ses voisins. Le score LOF est calculé en déterminant d'abord la densité d'accessibilité locale d'un point, qui est l'inverse de la distance moyenne entre ce point et ses voisins. Ensuite, le ratio de cette densité par rapport aux densités de ses voisins est calculé. Un point dont la densité est nettement inférieure à celle de ses voisins aura un score LOF élevé, indiquant ainsi une anomalie. Le modèle LOF a connu de nombreuses variantes, telles que COF [Tang et al., 2002], LOCI [Papadimitriou et al., 2003], CBLOF [He et al., 2003], LoOP [Kriegel et al., 2009], DILOF [Na et al., 2018], and CELOF [Chen et al., 2021a], qui améliorent ou modifient son concept fondamental pour répondre à des caractéristiques spécifiques des données ou à des défis particuliers de la tâche. En outre, tout comme KNN, LOF a été fondamental dans le développement d'algorithmes et de modèles plus sophistiqués [Pevný, 2016; Xu et al., 2022; Yong et al., 2022; Zou et al., 2023; Ding et al., 2024], ce qui souligne sa polyvalence et sa pertinence constante dans le domaine de la détection d'anomalies .

Malgré leur applicabilité générale, les méthodes basées sur la proximité sont confrontées à des défis, en particulier avec les données à haute dimension, où la malédiction de la dimensionnalité peut réduire l'efficacité des mesures de distance traditionnelles. Pour y remédier, des méthodes basées sur les angles, telles que *Angle-Based Outlier Detection* (ABOD) [Kriegel et al., 2008; Susto et al., 2017; Al-taei and Haeri, 2019; Taheri Sarteshnizi et al., 2024], ont été développées. ABOD évalue la variance des angles formés entre les vecteurs de différence d'un point et toutes les paires d'autres points dans l'ensemble de données. Cette méthode identifie les anomalies en se concentrant sur la distribution de ces angles au lieu des distances qui deviennent moins fiables dans les espaces à haute dimension. Les points ayant une variance élevée dans leurs angles sont considérés comme des anomalies, ce qui rend

ABOD particulièrement efficace pour détecter les anomalies dans les données à haute dimension, où les techniques traditionnelles à base de distance échouent.

Les méthodes basées sur la proximité offrent une approche claire et interprétable pour la détection d'anomalies, en exploitant la structure inhérente des données. Elles excellent dans les environnements où les clusters de données et les variations de densité sont prononcés. Cependant, leur applicabilité restreinte pour les données à haute dimension et leur sensibilité aux paramètres peuvent poser des défis, nécessitant des techniques avancées et un développement continu pour maintenir leur efficacité.

1.3.3.3 Approche basée sur la région

Les méthodes de détection d'anomalies basées sur la région ou le domaine consistent à établir une frontière autour de ce qui est considéré comme des données normales et à identifier tous les points qui se situent en dehors de cette frontière comme des anomalies [Pimentel et al., 2014; Solarz et al., 2017; Ruff et al., 2021]. Ces approches utilisent généralement des techniques telles que les machines à vecteurs de support pour définir le domaine des données normales dans l'espace des caractéristiques. Ces méthodes se concentrent sur la maximisation de la marge entre les points de données normaux et la frontière, ce qui leur permet de distinguer efficacement les anomalies des données normales.

Une technique fondamentale dans ce domaine est la machine à vecteurs de support à classe unique (*One-Class Support Vector Machine*, OCSVM) [Schölkopf et al., 2001; Li et al., 2003; Shinnou and Sasaki, 2010; Beghi et al., 2014; Solarz et al., 2017; Miao et al., 2019; Todkar et al., 2021; Yang et al., 2021; Avola et al., 2022]. L'OCSVM mappe les données d'entrée vers un espace de dimension supérieure à l'aide d'une fonction noyau, puis trouve un hyperplan à marge maximale qui sépare les points de données normaux de l'origine (le point de référence dans l'espace des caractéristiques où toutes les valeurs des caractéristiques sont nulles), isolant ainsi effectivement toutes les données normales d'un côté de l'hyperplan. Cet hyperplan définit la limite des données normales, et les points qui se trouvent en dehors de cette limite sont considérés comme des anomalies. De même, la description de données par vecteurs de support (*Support Vector Data Description*, SVDD) [Tax and Duin, 2004; Banerjee et al., 2007; Tao et al., 2007; Duong et al., 2015; Liu and Gryllias, 2020] est une autre technique clé qui construit une hypersphère autour des points de données normaux dans l'espace des caractéristiques. Le rayon et le centre de cette hypersphère sont optimisés pour inclure autant de points normaux que possible tout en minimisant le volume de la sphère. Les points qui se trouvent à l'extérieur de cette sphère sont signalés comme des anomalies.

Les progrès relativement récents impliquent l'intégration de réseaux neuronaux dans les modèles traditionnels, ce qui a conduit au développement des approches telles que DeepSVDD et DeepSAD. DeepSVDD [Ruff et al., 2018; Erfani et al., 2020; Zhang and Deng, 2021; Li et al., 2022a; Liu et al., 2024] étend le concept SVDD en utilisant des réseaux neuronaux pour extraire des caractéristiques mieux adaptées à la détection d'anomalies. Le réseau est entraîné à mapper les données normales vers une région avec un volume minimisé dans l'espace des caractéristiques, et les anomalies sont identifiées comme des points qui tombent en dehors de cette région. DeepSVDD a ensuite évolué vers *Deep Semi-Supervised Anomaly Detection* (DeepSAD) [Ruff et al., 2020], qui incorpore à la fois des données étiquetées et non

étiquetées pendant l'entraînement. Un autre modèle innovant, *Context Vector Data Description* (CVDD) [Ruff et al., 2019; Mu et al., 2021b], applique ces principes spécifiquement aux données textuelles, en analysant les contextes textuels pour identifier efficacement les documents ou phrases déviants.

Ces méthodes basées sur les régions offrent plusieurs avantages, notamment la flexibilité dans le traitement des diverses distributions de données et l'efficacité dans les contextes à haute dimension grâce à l'utilisation de fonctions noyau. L'adaptabilité des frontières à la forme de la distribution des données normales améliore encore leur utilité. Cependant, ces méthodes présentent également certaines contraintes. Elles sont sensibles au réglage spécifique des paramètres comme le type de noyau et le paramètre de régularisation, et leur complexité algorithmique peut être prohibitive pour de grands ensembles de données. En outre, leur performance peut être significativement affectée par des données bruyantes ou des classes chevauchantes, ce qui nécessite une préparation supplémentaire des données pour obtenir des résultats optimaux [Pimentel et al., 2014; Nassif et al., 2021].

1.3.3.4 Approche basée sur la reconstruction

La détection d'anomalies basée sur la reconstruction identifie les anomalies en reconstruisant les données à l'aide d'un modèle entraîné sur des données normales et en dérivant des scores à partir de l'erreur de reconstruction [Pimentel et al., 2014; An and Cho, 2015; Aggarwal, 2017a; Chen et al., 2018c; Zenati et al., 2018; Thudumu et al., 2020]. Cette approche repose sur l'hypothèse que les données normales peuvent être reconstruites exactement par un tel modèle, tandis que les anomalies ne peuvent pas être reconstruites aussi précisément. Les méthodes basées sur les sous-espaces et les réseaux neuronaux sont couramment utilisées dans ce cadre pour apprendre une représentation compressée des données normales.

L'une des techniques classiques basées sur la reconstruction est l'analyse en composantes principales (*Principal Component Analysis*, PCA) [Shyu et al., 2003, 2006; Brauckhoff et al., 2009; Pascoal et al., 2012; Ding and Kolaczyk, 2013; Kumar and Ravi, 2017; Hoang and Nguyen, 2018]. La PCA projette les données sur un espace de dimension inférieure défini par les composantes principales qui capturent la plus grande variance dans les données. Les anomalies sont identifiées en mesurant l'erreur de reconstruction, à savoir la différence entre les données originales et leur projection sur les composantes principales. Outre la PCA standard, différentes variantes de PCA ont été conçues pour répondre à des défis spécifiques. Par exemple, la PCA à noyaux (*Kernel Principal Component Analysis*, KPCA) [Hoffmann, 2007; Xiao et al., 2014; Yong et al., 2017; Wang et al., 2021; Pan et al., 2022; Simmini et al., 2022] étend cette approche pour traiter les relations non linéaires dans les données en utilisant des fonctions de noyau.

Les réseaux de neurones, notamment les auto-encodeurs et les auto-encodeurs variationnels (*Variational AutoEncoder*, VAE), sont également largement utilisés dans ce cadre. Les auto-encodeurs se composent d'un encodeur et d'un décodeur, où l'encodeur mappe les données d'entrée vers un espace latent, et le décodeur reconstruit l'entrée à partir de cette représentation latente. Des erreurs de reconstruction élevées indiquent des anomalies [Sakurada and Yairi, 2014; Chen et al., 2017; Zhou and Paffenroth, 2017; Chen et al., 2018c; Kieu et al., 2019; Sarvari et al., 2021; Abhaya and Patra, 2023]. D'autre part, les VAE introduisent une approche probabiliste en

modélisant les données d'entrée comme une distribution et en cherchant à régénérer l'entrée à partir de cette distribution, fournissant ainsi une mesure d'incertitude dans les reconstructions, ce qui est utile pour la détection d'anomalies.

Les réseaux antagonistes génératifs (*Generative Adversarial Networks*, GANs) [Goodfellow et al., 2014] ont également été adaptés pour la détection d'anomalies. Dans le cadre standard des GANs, un générateur crée des échantillons de données, et un discriminateur tente de distinguer entre les échantillons réels et générés. Pour la détection d'anomalies, le générateur est entraîné à reconstruire des échantillons de données normales, et le discriminateur est utilisé pour évaluer la qualité de ces reconstructions. L'erreur de reconstruction, à savoir la différence entre l'entrée et la sortie générée, est utilisée comme un score d'anomalie [Beula Rani and Sumathi M. E, 2020; Di Mattia et al., 2021; Xia et al., 2022; Li and Li, 2023]. AnoGAN [Schlegl et al., 2017, 2019; Shin et al., 2020; Kim et al., 2021a; Singh and Reddy, 2024] entraîne un GAN sur des données normales et utilise le générateur pour trouver la représentation la plus proche d'un échantillon de test dans l'espace latent, minimisant l'erreur de reconstruction pour les données normales et en identifiant les anomalies par leurs erreurs plus élevées. Ganomaly [Akcay et al., 2019; Jiang et al., 2019; Luo et al., 2021; Liu et al., 2022a] améliore AnoGAN en introduisant une architecture encodeur-décodeur-encodeur, où un second encodeur mappe l'échantillon généré de nouveau dans l'espace latent, apprenant ainsi des caractéristiques plus robustes pour détecter les anomalies. *Adversarially Learned Anomaly Detection* (ALAD) [Zenati et al., 2018; Knapp et al., 2021; Zhang and Zuo, 2021] améliore les GANs standards en employant des GANs bidirectionnels (BiGANs) [Donahue et al., 2017], qui assurent la cohérence entre les données d'entrée et les espaces latents. Il utilise des pertes de cohérence du cycle et deux discriminateurs pour améliorer l'exactitude de la détection en comparant les erreurs de reconstruction à la fois dans les données et les espaces latents.

Les méthodes basées sur la reconstruction peuvent traiter des structures de données complexes et à haute dimension. Elles sont particulièrement efficaces pour les contextes d'apprentissage non supervisés et semi-supervisés. Cependant, elles exigent d'importantes ressources informatiques, en particulier lors de l'utilisation de modèles profonds sur de grands ensembles de données. De plus, la complexité et le manque d'interprétabilité de ces modèles peuvent les rendre difficiles à comprendre et à diagnostiquer dans des applications pratiques.

1.3.3.5 Approche basée sur la théorie de l'information

La détection d'anomalies basée sur la théorie de l'information utilise des métriques de la théorie de l'information, telles que l'entropie, l'information mutuelle et les mesures de divergence, pour identifier les anomalies au sein des ensembles de données. Ces méthodes reposent sur l'hypothèse fondamentale que les données normales suivent une distribution statistique spécifique, et les anomalies perturbent cette distribution en modifiant le contenu informationnel des données, ce qui entraîne des changements détectables dans les mesures théoriques de l'information [Chandola et al., 2009; Pimentel et al., 2014; Ahmed et al., 2016b].

Au cœur de ces méthodes est le calcul du contenu informationnel à l'aide de mesures telles que l'entropie et l'entropie relative. L'entropie, par exemple, mesure le niveau d'incertitude ou degré d'aléa au sein des données. Les données normales pré-

sentent généralement une entropie plus faible en raison de leurs patterns prévisibles, tandis que les données anormales introduisent de l'imprévisibilité, augmentant ainsi l'entropie. En comparant l'entropie de l'ensemble de données avant et après le retrait de certains points de données, les anomalies peuvent être détectées comme les points qui augmentent significativement l'entropie lorsqu'ils sont inclus [Wagner and Plattner, 2005; Jiang et al., 2010; Müter and Asaj, 2011; Navaz et al., 2013; Bereziński et al., 2015; Howedi et al., 2020].

Une autre mesure critique utilisée dans ce cadre est la divergence de Kullback-Leibler (KLD), également connue sous le nom d'entropie relative, qui quantifie la différence entre les distributions de probabilité des données normales et des données observées. Cette méthode suppose que les données normales suivent une distribution de probabilité connue, et les anomalies dévient de cette distribution, entraînant des valeurs de divergence importantes. En calculant la KLD entre les distributions des données normales et observées, les anomalies peuvent être identifiées comme les points causant des déviations significatives [Anderson and Haas, 2011; Zeng et al., 2014; Youssef et al., 2016; Xie et al., 2017; Huang et al., 2021].

L'information mutuelle mesure la dépendance entre les variables au sein de l'ensemble de données. Dans les données normales, les variables présentent de fortes dépendances, quantifiées par des valeurs élevées d'information mutuelle. Les anomalies perturbent ces dépendances, ce qui entraîne une réduction de l'information mutuelle. En mesurant l'information mutuelle entre les variables dans l'ensemble de données, les anomalies peuvent être détectées comme les points qui affaiblissent les dépendances normales [Kopylova et al., 2008; Amiri et al., 2011; Bian et al., 2019; Ye et al., 2021; Hu et al., 2023].

Ces métriques de la théorie de l'information sont souvent employées aux côtés de techniques de compression telles que la PCA [Kanda et al., 2013; Callegari et al., 2014; Harmouche et al., 2014, 2015; Hong et al., 2016; Wang et al., 2016b; Bounoua et al., 2020]. L'idée est que les données normales, en raison de leurs patterns réguliers, se compressent efficacement, tandis que les données anormales, qui perturbent ces patterns, entraînent une compression moins efficace. En examinant les différences de taux de compression, les anomalies peuvent être identifiées comme les points qui augmentent significativement la taille de la compression.

Les méthodes à base de théorie de l'information ne dépendent pas de modèles ou d'hypothèses spécifiques sur la distribution des données, ce qui les rend polyvalentes et applicables à divers types de données. Elles peuvent traiter efficacement des données à haute dimension, car elles se concentrent sur le contenu informationnel plutôt que sur la dimensionnalité. De plus, ces techniques sont sensibles aux changements dans la distribution des données, ce qui les rend efficaces pour détecter des anomalies subtiles qui pourraient être ratées par d'autres méthodes. Malgré ces avantages, les méthodes basées sur la théorie de l'information sont confrontées à des défis. Elles exigent des calculs considérables, surtout lorsqu'il s'agit de grands ensembles de données à haute dimension. De plus, ces méthodes sont sensibles aux paramètres, ce qui nécessite un réglage attentif pour garantir l'exactitude. Les résultats de ces approches peuvent parfois être difficiles à interpréter, car ils fournissent une mesure du contenu informationnel plutôt qu'une indication directe de ce qui constitue une anomalie.

1.3.3.6 Approche basée sur l'apprentissage ensembliste

Les méthodes ensemblistes constituent l'une des tendances majeures dans la recherche récente en détection d'anomalies [Xu et al., 2019b; Villa-Pérez et al., 2021; Han et al., 2022]. Elles combinent plusieurs modèles de détection d'anomalies pour améliorer les performances globales de détection. Cette approche repose sur le concept de l'apprentissage ensembliste, qui suppose que la combinaison des forces de plusieurs modèles peut aboutir à un ensemble plus performant, plus robuste et plus généralisable qu'un modèle unique [Zhou, 2012; Nun et al., 2016; Dong et al., 2020; Zhou and Zhou, 2021]. En profitant des points forts des modèles individuels, les méthodes ensemblistes peuvent capturer divers aspects des anomalies plus efficacement, fournissant ainsi des résultats plus plausibles.

L'une des techniques principales dans ce cadre est le *Bagging*, ou *Bootstrap Aggregating* [Zimek et al., 2013; Pasillas-Díaz and Ratté, 2017; Biswas and Samanta, 2021; Ouyang et al., 2021; Reddy et al., 2021]. Cette approche consiste à générer plusieurs sous-ensembles de l'ensemble de données par bootstrap et à entraîner un modèle sur chaque sous-ensemble. Les prédictions de ces modèles sont ensuite agrégées, généralement par moyenne ou vote, pour former une décision finale sur les points de données. Un exemple notable de cette approche est le modèle *Isolation Forest* (IForest) [Liu et al., 2008, 2012; Ding and Fei, 2013; Xu et al., 2017], qui construit un ensemble d'arbres pour isoler les anomalies en partitionnant récursivement les données. Son adaptation, *Deep Isolation Forest* (DIF) [Xu et al., 2023b], étend ce concept en utilisant des techniques d'apprentissage profond pour traiter des structures de données plus complexes.

Une autre technique, le *Boosting* [Salehi et al., 2016; Sundqvist et al., 2020; Xing and Liu, 2020; Ikram et al., 2021; Shahzad et al., 2022], se concentre sur la construction séquentielle de modèles, chacun visant à corriger les erreurs de ses prédécesseurs, améliorant ainsi l'exactitude globale des prédictions. Cette approche garantit que les points de données difficiles à prédire, qui sont souvent des anomalies, reçoivent plus d'attention dans les modèles suivants. Un exemple typique est *Extreme Gradient Boosting for Outlier Detection* (XGBOD) [Zhao and Hryniewicki, 2018], qui combine le gradient boosting [Chen and Guestrin, 2016] et les méthodes de détection d'anomalies pour corriger séquentiellement les erreurs et ajuster les poids des modèles en fonction de leurs performances, aboutissant à un classificateur performant en détection d'anomalies.

Le *Stacking* [Ouyang et al., 2018; Fatemifar et al., 2020; Nkenyereye et al., 2020; Zhang et al., 2021b] complète ces approches en intégrant les prédictions de plusieurs modèles à l'aide d'un méta-modèle. L'hypothèse sous-jacente est que l'architecture du modèle en couches capture un ensemble plus riche de complexités des données, améliorant ainsi les capacités de détection. Cette stratégie entraîne plusieurs modèles de base indépendamment, puis utilise leurs prédictions comme entrées pour un méta-modèle afin de prendre la décision finale concernant les anomalies.

Le mélange d'experts (*Mixture of Experts*, MoE) [Nun et al., 2016; Pham et al., 2021; Yu et al., 2021; Schulze et al., 2022; Zhao et al., 2023] est l'une des approches les plus remarquables de l'apprentissage ensembliste profond. Il combine plusieurs réseaux spécialisés, appelés experts, pour résoudre des problèmes comme la détection d'anomalies. Dans un modèle MoE, chaque expert se spécialise dans une partie particulière des données d'entrée, et un réseau de contrôle décide quel expert doit

être utilisé pour chaque point de donnée. Le réseau de contrôle apprend à sélectionner l'expert le plus adapté en fonction des caractéristiques du point donnée, ce qui permet d'appliquer le meilleur modèle pour chaque cas. Cette méthode améliore les performances globales en combinant les forces de modèles spécialisés, chacun étant compétent pour détecter les anomalies dans des régions spécifiques des données.

Les méthodes ensemblistes offrent plusieurs avantages. Elles améliorent l'exactitude et la robustesse de détection par rapport aux modèles individuels, traitent divers types d'anomalies et de distributions de données, et réduisent le risque de surapprentissage en exploitant plusieurs modèles. Cependant, elles présentent également des limitations. Ces méthodes augmentent la complexité de calcul et les exigences en termes de ressources informatiques, nécessitent une sélection attentive des modèles de base et de la stratégie d'apprentissage ensembliste, et peuvent rendre l'interprétation des résultats plus complexe en raison de la combinaison de plusieurs modèles.

1.3.3.7 Approche basée sur l'apprentissage de scores d'anomalie

L'apprentissage de scores de bout en bout représente un autre changement important dans la détection d'anomalies, évoluant des approches traditionnelles en deux étapes, à savoir l'établissement de la norme et l'évaluation des écarts, vers une méthode plus intégrée grâce à l'apprentissage profond [Pang et al., 2021b; Xu et al., 2019b; Han et al., 2022; Jiang et al., 2023]. Cette approche optimise la détection d'anomalies en apprenant les scores d'anomalie directement à partir des données dans un processus continu et unique, évitant ainsi les inconvénients de la séparation entre l'apprentissage des caractéristiques et le calcul des scores.

De nombreuses méthodes ont été développées dans ce cadre. *Deep Anomaly Detection with Deviation Networks* (DevNet) [Pang et al., 2019], par exemple, optimise une fonction de perte de déviation à base de score z pour aligner les scores d'anomalie prédits avec une distribution gaussienne a priori, améliorant ainsi la capacité à détecter des anomalies dans les données de haute dimension. Une autre méthode, *Pairwise Relation prediction Network* (PreNet) [Pang et al., 2020], se concentre sur l'apprentissage des relations par paire entre les points de données. En prédisant les scores d'anomalie relatifs entre les paires d'échantillons, PreNet identifie efficacement les instances anormales à l'aide d'une approche comparative, en exploitant les relations contextuelles au sein des données. Par ailleurs, *Adversarially Learned One-Class* (ALCCO) [Sabokrou et al., 2018] s'appuie sur l'apprentissage antagoniste et la classification à classe unique pour entraîner un classificateur à distinguer les instances normales et anormales de manière intégrée.

Les avantages de l'apprentissage de scores de bout en bout sont multiples. Il est particulièrement efficace pour traiter des données à haute dimension dans un cadre unifié qui intègre l'extraction des caractéristiques, l'évaluation des scores d'anomalie et la prise de décision, minimisant ainsi les risques d'erreurs inhérents aux processus multi-étapes [Pang et al., 2019; Hamdi et al., 2021]. De plus, cette approche permet d'incorporer des modèles pré-entraînés et des techniques d'apprentissage par transfert, améliorant ainsi les performances sur des tâches spécifiques. Cependant, il existe également des limites notables. L'entraînement des modèles profonds exige des ressources informatiques considérables. De plus, les opérations des modèles sont souvent opaques, rendant difficile l'interprétation des raisons derrière leurs décisions. Ce manque de transparence peut constituer un inconvénient majeur, notamment dans

les applications critiques où comprendre le processus de décision est essentiel.

1.4 Évaluation

L'évaluation efficace des algorithmes est une étape cruciale dans le développement des systèmes de détection d'anomalies. Néanmoins, comme les anomalies sont rares par nature et que le processus d'annotation peut être très subjectif, il s'avère souvent difficile d'obtenir suffisamment d'étiquettes de référence pour une évaluation complète. Par conséquent, les applications dans le monde réel ont tendance à s'appuyer plutôt sur des études de cas avec une analyse qualitative et des techniques de visualisation que sur des évaluations quantitatives systématiques pour évaluer les performances des systèmes de détection d'anomalies.

En ce qui concerne l'évaluation quantitative, les études antérieures ont parfois utilisé des mesures de validité interne telles que le score de Silhouette [Rousseeuw, 1987] ou l'indice de Davies-Bouldin [Davies and Bouldin, 1979]. Ces mesures évaluent la qualité de la détection sur la base des propriétés intrinsèques des données, sans avoir recours à des connaissances externes ou à des étiquettes de référence. Ils donnent un aperçu de la structure de l'ensemble de données et aident à perfectionner les modèles en évaluant les caractéristiques de clustering, contribuant ainsi à la détection d'anomalies.

Toutefois, ces mesures de validité interne, principalement conçues pour les tâches de clustering, sont intrinsèquement limitées pour l'évaluation de la détection d'anomalies. En matière de clustering, l'objectif est de maximiser la similarité à l'intérieur d'un cluster et de minimiser la similarité entre les clusters. En revanche, la détection d'anomalies vise à identifier les points de données qui s'écartent de manière significative de la majorité. Ainsi, ces mesures offrent seulement des comparaisons relatives et ne mesurent pas directement l'efficacité d'un algorithme à isoler les événements rares. De plus, elles reposent souvent sur l'hypothèse que les données peuvent naturellement être segmentées en clusters distincts, une hypothèse qui n'est pas nécessairement appropriée pour la détection d'anomalies. En outre, les anomalies peuvent perturber les mesures telles que le score de Silhouette ou l'indice de Davies-Bouldin en faussant les calculs de distance. Cette distorsion donne l'impression que les clusters sont plus dispersés et moins compacts, ce qui conduit à des évaluations inexactes.

Au vu de ces limites, les mesures de validité interne ne sont pas couramment utilisées dans la détection d'anomalies. À la place, les chercheurs se tournent vers des mesures de validité externe, qui évaluent les performances d'un algorithme par rapport à une vérité de référence fournie de l'extérieur. Le plus souvent, il s'agit d'utiliser des benchmarks externes ou des jeux de données synthétiques qui comportent des anomalies déjà connues. Ces ensembles de données proviennent généralement de tâches de classification (déséquilibrée), où des étiquettes rares servent de substituts aux anomalies réelles. Par conséquent, la détection d'anomalies est généralement évaluée à l'aide de métriques appropriées pour la classification binaire, en particulier dans des contextes où les données sont extrêmement déséquilibrées. Les mesures couramment utilisées en classification binaire, telles que la matrice de confusion, la précision, le rappel, le F-score, l'AUCROC et l'AUCPR, sont également privilégiées pour évaluer la détection d'anomalies.

Comme nous l'avons indiqué dans la section §1.2, la sortie d'un système de dé-

tection d'anomalies peut prendre deux formes : une étiquette binaire ou un score d'anomalie continu. La plupart des systèmes produisent d'abord un score d'anomalie continu, qui est ensuite converti en une étiquette binaire à l'aide d'un seuil spécifique. Les mesures d'évaluation varient selon le type de sortie : par exemple, la précision et le rappel ne conviennent que pour les étiquettes discrètes, typiques de la classification binaire, tandis que la courbe ROC exige une probabilité ou un score d'anomalie continu. Dans notre discussion sur les métriques d'évaluation, nous traiterons les étiquettes binaires comme résultant de l'application d'un seuil à un score d'anomalie continu.

1.4.1 Matrice de confusion

Considérons une tâche de détection d'anomalies sur un ensemble de données $\mathcal{D}$. L'ensemble des anomalies de vérité terrain est désigné par V_a. Pour un seuil donné s, l'ensemble des anomalies prédites est désigné par $A(s)$. La taille de $A(s)$ change en fonction de s, typiquement en diminuant lorsque le seuil se lève. Les résultats du système de détection peuvent être classés en quatre catégories de base :

— **Vrais Positifs (VP)** : Les anomalies *correctement* identifiées comme des anomalies ($A(s) \cap V_a$).
— **Vrais Négatifs (VN)** : Les instances normales *correctement* identifiées comme normales ($\overline{A(s)} \cap \overline{V_a}$).
— **Faux Positifs (FP)** : Les instances normales *incorrectement* identifiées comme des anomalies ($A(s) \cap \overline{V_a}$).
— **Faux Négatifs (FN)** : Les anomalies *incorrectement* identifiées comme normales ($\overline{A(s)} \cap V_a$).

Ces quatre résultats peuvent être organisés sous la forme d'un tableau de contingence 2×2, où les lignes correspondent à la vérité de référence (condition réelle) et les colonnes à la prédiction (condition prédite). Ce tableau est appelé **matrice de confusion** ou **matrice d'erreur**. Il fournit une visualisation directe des performances de l'algorithme de détection d'anomalies et sert souvent de métrique d'évaluation la plus élémentaire pour cette tâche.

1.4.2 Ratios importants

À partir des quatre résultats de base, on peut dériver plusieurs ratios qui servent d'indicateurs importants pour la performance des systèmes de détection d'anomalies, tant dans la recherche que dans les applications industrielles.

1.4.2.1 Taux de faux négatifs

Le taux de faux négatifs (TFN) est défini comme la proportion de véritables anomalies que l'algorithme classe à tort comme normales :

$$\text{TFN} = \frac{\text{FN}}{\text{FN} + \text{VP}} = \frac{|\overline{A(s)} \cap V_a|}{|V_a|}$$

Il mesure la probabilité que de véritables anomalies ne soient pas détectées par le système, et est également connu sous le nom de **taux de ratés**. Le TFN est particulièrement utile lorsque le défaut de détection d'une anomalie peut entraîner de

graves conséquences. Par exemple, dans le domaine de la veille stratégique, omettre un événement important concernant l'évolution stratégique d'un concurrent peut se traduire par des pertes considérables pour l'entreprise. Dans ce cas, il convient d'augmenter le seuil pour minimiser le taux de ratés.

1.4.2.2 Taux de faux positifs

Le taux de faux positifs (TFP) est la proportion d'instances normales que l'algorithme classe à tort comme des anomalies :

$$\text{TFP} = \frac{\text{FP}}{\text{VN} + \text{FP}} = \frac{|A(s) \cap \overline{V_a}|}{|\overline{V_a}|}$$

Cet indicateur de performance est important dans les applications pratiques, car il mesure la probabilité que le système déclenche de **fausses alarmes**. La TFP est particulièrement utile dans les scénarios où le coût du traitement des fausses alarmes est élevé. Par exemple, dans le cadre de la surveillance stratégique, les signaux faibles faussement détectés peuvent gaspiller des ressources substantielles en investigations et analyses inutiles. Il est donc préférable de baisser le taux de fausses alarmes en élevant le seuil.

1.4.2.3 Précision

La précision, également connue sous le nom de valeur prédictive positive (VPP), est la proportion d'instances classées comme des anomalies qui le sont réellement :

$$\text{Précision} = \frac{\text{VP}}{\text{VP} + \text{FP}} = \frac{|A(s) \cap V_a|}{|A(s)|}$$

Elle mesure l'exactitude des prédictions positives faites par l'algorithme. Cette mesure est simple et répond directement à la question : « Quand le modèle prédit une anomalie, quelle est la probabilité qu'il soit correct ? »

1.4.2.4 Rappel

Le rappel, également connu sous le nom de sensibilité ou de taux de vrais positifs (TVP), est la proportion de véritables anomalies correctement identifiées par l'algorithme :

$$\text{Rappel} = \frac{\text{VP}}{\text{VP} + \text{FN}} = \frac{|A(s) \cap V_a|}{|V|}$$

Il mesure le degré d'exhaustivité des prédictions positives et répond à la question : « Combien d'anomalies réelles le modèle a-t-il détectées ? »

1.4.2.5 Précision et rappel vs. FPR et FNR

La précision et le rappel sont directement liés aux objectifs principaux de la tâche de détection d'anomalies : identifier autant d'anomalies réelles que possible (rappel) tout en maintenant un faible nombre de fausses alertes (précision). Ils offrent une évaluation claire et intuitive de la performance d'un modèle en se concentrant sur la qualité de ses prédictions.

En revanche, le taux de faux positifs (FPR) et le taux de faux négatifs (FNR) offrent une compréhension plus détaillée des erreurs mais sont souvent moins alignés

avec les objectifs directs de la détection d'anomalies. Ils sont généralement utilisés dans des analyses diagnostiques plus détaillées plutôt qu'en tant que métriques de performance principales.

1.4.3 F-score

Bien que la précision et le rappel fournissent un aperçu utile de la performance des modèles de détection d'anomalies, il est souvent nécessaire de les considérer conjointement pour obtenir une compréhension globale. Pour y parvenir, le F-score, ou F-measure, est proposé. L'idée générale du F-score est de fournir une mesure unique qui permette un compromis entre la précision et le rappel, donnant ainsi une évaluation complète de l'efficacité d'un modèle.

1.4.3.1 F_1-score

La forme de base du F-score, le F_1-score , est défini comme la moyenne harmonique de la précision et du rappel :

$$F_1 = 2 \times \frac{\text{Précision} \times \text{Rappel}}{\text{Précision} + \text{Rappel}}$$

En équilibrant la précision et le rappel, le F_1-score évite les inconvénients de l'utilisation isolée de ces mesures et fournit une évaluation plus complète de la performance du modèle. Le F_1-score est particulièrement efficace dans les scénarios où l'équilibre entre la précision et le rappel est crucial.

1.4.3.2 F_β-score

Toutefois, dans certains contextes, surtout lorsque le coût associés aux faux négatifs et aux faux positifs n'est pas le même, la pondération égale du F_1-score pourrait ne paraitre pas idéale. Par conséquent, le F_1-score est étendu à une forme plus générale, le F_β-score :

$$F_\beta = (1 + \beta^2) \times \frac{\text{Précision} \times \text{Rappel}}{\beta^2 \times \text{Précision} + \text{Rappel}}$$

où β est un paramètre qui détermine le poids du rappel dans le score combiné. Deux valeurs couramment utilisées pour β sont 2, qui donne plus de poids au rappel qu'à la précision, et 0.5, ce qui donne plus de poids à la précision qu'au rappel. Le score F_β est idéal pour les applications où l'importance de la précision et du rappel diffère. Par exemple, dans le diagnostic médical, $\beta > 1$ (F_2, F_3, etc.) met l'accent sur le rappel, ce qui est crucial lorsque manquer une anomalie (maladie) est plus grave qu'une fausse alerte. En utilisant ces F-scores, les utilisateurs peuvent adapter leur approche d'évaluation aux exigences spécifiques de leurs tâches de détection d'anomalies, en s'assurant que la métrique choisie s'aligne avec les priorités opérationnelles et commerciales.

1.4.4 PRC et AUCPR

Alors que le F-score fournisse une métrique équilibrée pour évaluer la performance d'un modèle de détection d'anomalies, il ne parvient pas toujours à donner une vue globale des capacités d'un modèle, en particulier en fonction de différents seuils de décision. Pour y remédier, nous nous tournons vers la courbe précision-rappel (PRC) et son aire sous la courbe (AUCPR).

1.4.4.1 Courbe de Précision-Rappel (PRC)

En changeant le seuil s, il est possible de dresser une courbe en opposant la précision au rappel. Cette courbe est connue comme la courbe précision-rappel (*Precision-Recall Curve*, PRC). La PRC est une représentation graphique qui illustre le compromis entre la précision et le rappel à différents seuils d'un modèle de détection d'anomalies. Chaque point de la courbe correspond à un seuil de décision spécifique, indiquant la précision et le rappel obtenus par le modèle à ce seuil. Contrairement au F-score, qui fournit une métrique unique pour l'évaluation d'un algorithme à un seuil spécifique, la PRC offre une méthode visuelle et intuitive pour observer comment la précision et le rappel varient en fonction des changements du seuil de décision. Cela permet une analyse plus approfondie des performances du modèle pour différents seuils. En examinant la courbe, les utilisateurs peuvent identifier le seuil qui offre des meilleurs résultats pour les besoins de leur application spécifique.

1.4.4.2 Aire sous la courbe PRC (AUCPR)

L'aire sous la courbe précision-rappel (AUCPR) est une valeur scalaire unique qui résume la performance du modèle pour tous les seuils. L'AUCPR est calculé en intégrant la surface sous la courbe de précision-rappel, avec une valeur comprise entre 0 et 1 :

— Un score élevé indique que le modèle est performant à travers une variété de seuils, en maintenant un équilibre élevé entre la précision et le rappel.

— Un score de 1 indique une précision et un rappel parfaits, ce qui est idéal.

— Un score de 0,5 suggère une performance au niveau de prédiction aléatoire.

— Les scores entre 0,5 et 1 correspondent à des degrés variables de performance du modèle en termes de précision et de rappel.

1.4.5 ROC et AUCROC

Alors que la PRC et l'AUCPR offrent des perspectives intéressantes sur le compromis entre la précision et le rappel, un autre outil indispensable pour évaluer la performance du modèle à travers différents seuils est la courbe ROC (*Receiver Operating Characteristic*) et son aire sous la courbe correspondante (AUCROC).

1.4.5.1 *Receiver Operating Characteristic* (ROC)

La courbe ROC (*Receiver Operating Characteristic*) est une représentation graphique qui illustre la capacité d'un modèle à distinguer entre les classes normales et anormales à travers tous les seuils de détection possibles. Cette courbe trace le taux de vrais positifs (sensibilité ou rappel) contre le taux de faux positifs ($1 -$ spécificité) pour chaque seuil. Plus la courbe ROC est proche du coin supérieur gauche du graphique, meilleure est la performance du modèle. Un modèle parfait aurait une courbe qui atteint le coin supérieur gauche, indiquant un TVP de 1 (100% de sensibilité) et un FPR de 0. La courbe montre comment les compromis entre les gains (vrais positifs) et les coûts (faux positifs) changent lorsque le seuil de décision varie. Elle permet de sélectionner un seuil qui équilibre la sensibilité et la spécificité en fonction des besoins d'une application spécifique.

1.4.5.2 Aire sous la courbe ROC (AUCROC)

L'aire sous la courbe ROC (AUCROC) est une mesure résumée qui capture la performance globale du modèle sur tous les seuils. Une valeur de l'AUCROC s'étend de 0 à 1, où 1 indique une performance parfaite et 0,5 représente une performance équivalente à une prédiction aléatoire. Cette métrique est particulièrement avantageuse car elle fournit un seul chiffre récapitulatif permettant de comparer objectivement différents modèles, indépendamment de tout seuil spécifique.

1.5 Données et applications

La détection d'anomalies est appliquée dans de nombreux domaines, impliquant une variété de types de données. Cette section se penche sur les applications de la détection d'anomalies, organisées selon la nature des données concernées.

1.5.1 Séries Temporelles

Omniprésentes dans les activités humaines, les séries temporelles consistent en des séquences de points de données enregistrés à intervalles réguliers, chacun marqué par un horodatage [Chandola et al., 2009; Shaukat et al., 2021; Schmidl et al., 2022]. Ces données sont caractérisées par leur ordre temporel, ce qui nécessite une analyse de la dépendance temporelle des points de données. Cette dépendance temporelle est cruciale car elle influence l'identification des tendances et des variations saisonnières, qui ne sont pas pertinentes pour d'autres types de données [Aminikhanghahi and Cook, 2017; Shaukat et al., 2021]. Les séries temporelles constituent l'un des types de données les plus étudiés en détection d'anomalies, avec des applications étendues dans divers domaines. En finance, elles permettent de détecter les fraudes par carte de crédit [Devaki et al., 2014; Tran et al., 2018; Moschini et al., 2021; Zhang et al., 2021a; Iqbal and Amin, 2024] et d'identifier les comportements manipulateurs sur les marchés boursiers [Luo et al., 2008; Cao et al., 2015; Golmohammadi and Zaiane, 2015, 2017; Tallboys et al., 2022]. Le secteur industriel les utilise pour la maintenance prévisionnelle [De Benedetti et al., 2018; Makridis et al., 2020; Carrasco et al., 2021; Coelho et al., 2022; Raparthi, 2023] et la surveillance de la consommation d'énergie [Chou and Telaga, 2014; Janetzko et al., 2014; Chahla et al., 2020; Liu et al., 2021c], ce qui est crucial pour maintenir l'efficacité opérationnelle. Dans le domaine de la santé, la détection d'anomalies dans les séries temporelles est essentielle pour le suivi des pandémies [Homayouni et al., 2021; Jombart et al., 2021; Kumar et al., 2021] et la surveillance continue de la santé des patients [Chauhan and Vig, 2015; Chuah and Fu, 2007; Presbitero et al., 2017; Salem et al., 2021; Sivapalan et al., 2022], ce qui améliore les pratiques médicales préventives et réactives. En outre, dans le secteur environnemental, elles contribuent à la surveillance du changement climatique [Das and Parthasarathy, 2009; Srinivasan et al., 2020; Wahyono et al., 2020; Li and Jung, 2021], fournissant des données cruciales pour la prise de décisions écologiques.

Les défis principaux de la détection d'anomalies dans les séries temporelles incluent la capture et la modélisation des dépendances temporelles pour identifier exactement les anomalies. Ces données contiennent souvent une saisonnalité et des tendances inhérentes, qui doivent être distinguées des anomalies réelles. De plus, les besoins de traitement en temps réel dans de nombreuses applications, telles que la

détection de fraudes financières ou la surveillance de la santé, exige des algorithmes capables de fonctionner rapidement et efficacement pour fournir des alertes en temps opportun.

1.5.2 Données spatiales

Les données spatiales, ou données géospatiales, concernent les informations liées à la localisation spécifique sur la surface de la Terre, qui englobent les coordonnées géographiques et les relations spatiales [Shekhar et al., 2003; Chen et al., 2008]. Ces données sont souvent associées à des éléments temporels pour analyser les changements au fil du temps, ce qui est connu sous le nom d'analyse spatio-temporelle [Atluri et al., 2018]. La détection d'anomalies est un aspect fondamental de l'analyse spatio-temporelle, avec de nombreuses applications dans divers secteurs. Dans la sécurité publique, elle joue un rôle crucial dans l'identification des épidémies [Karadayi et al., 2020; Iacus et al., 2021], la réponse aux catastrophes naturelles [Refonaa et al., 2015; Marzuoli and Liu, 2019] et la surveillance de l'état des infrastructures [De Melo Borges et al., 2016; Shu et al., 2023]. La planification urbaine utilise la détection d'anomalies spatiales pour gérer l'utilisation des terres [Qian et al., 2010; Kong et al., 2020] et surveiller les flux de trafic [Refonaa et al., 2015; Zhang et al., 2016], tandis que le secteur agricole en profite pour surveiller l'état des cultures [Moso et al., 2021; Dorbu and Hashemi-Beni, 2023] et analyser l'humidité du sol [Cuina et al., 2019; Greifeneder et al., 2019]. Ses applications dans la gestion environnementale incluent la détection de la déforestation [Hamunyela et al., 2016, 2020] et la surveillance de la qualité de l'air [Chen et al., 2018a; Lin et al., 2022].

Les défis de la détection des anomalies spatiales proviennent de la dépendance contextuelle et de la dynamique temporelle des données spatiales. Les anomalies spatiales sont souvent dépendantes du contexte. Ce qui est considéré comme une anomalie dans une région ou un contexte peut être normal dans un autre. Les méthodes de détection d'anomalies doivent donc incorporer des informations contextuelles pour identifier correctement les anomalies. De plus, de nombreux ensembles de données spatiales sont dépendants du temps, ce qui ajoute une couche supplémentaire de complexité. Les anomalies peuvent être temporelles aussi bien que spatiales, nécessitant des méthodes capables de gérer les données spatio-temporelles.

1.5.3 Images

Les images sont des représentations numériques d'informations visuelles enregistrées sous forme de matrices de pixels. Comme ce type de données est omniprésent dans l'industrie ainsi que dans la vie quotidienne, la détection d'anomalies dans les images trouve de nombreux scénarios d'application dans divers domaines. Dans le domaine de la santé, elle est largement utilisée pour le diagnostic médical [Xu et al., 2019a; Baur et al., 2021; Han et al., 2021a; van Hespen et al., 2021; Emin Sahin, 2022], où l'analyse des images médicales telles que les radiographies, IRM ou scanographie aide à détecter des anomalies comme les tumeurs. Dans le secteur manufacturier, la détection d'anomalies visuelles assure le contrôle de la qualité sur les chaînes de production [Tonnaer et al., 2019; Lyu and Manoochehri, 2021; Napoletano et al., 2021] et inspecte les cartes de circuits imprimés pour détecter les défauts [Volkau et al., 2019; Shi et al., 2020; Kim et al., 2021b]. En ce qui concerne les applications plus récentes, les véhicules autonomes s'appuient sur la détection d'anomalies

visuelles pour des tâches essentielles telles que la détection d'objets [Di Biase et al., 2021; Walambe et al., 2021; Kapusi et al., 2022] et l'alerte de franchissement involontaire de ligne [Han et al., 2021b].

Les défis posés par le traitement des dans le cadre de la détection d'anomalies incluent la haute dimensionnalité et la complexité des données visuelles, ce qui nécessite des ressources computationnelles importantes. Par ailleurs, les anomalies dans les images peuvent varier largement, exigeant des algorithmes capables de détecter des irrégularités subtiles ainsi que des distorsions plus prononcées.

1.5.4 Vidéos

Les vidéos, une séquence d'images ou de trames englobant les dimensions à la fois spatiales et temporelles, offrent une représentation dynamique des scènes au fil du temps. Cette combinaison d'informations temporelles et spatiales présente des défis uniques, car les vidéos héritent des complexités des images et des séries temporelles. La détection d'anomalies visuelles dynamiques trouve des applications dans de nombreux domaines. Dans la sécurité et la surveillance, elle est utilisée pour détecter des intrusions [Khaleghi and Moin, 2018; Nayak et al., 2019; Wang et al., 2023b] et identifier des comportements suspects [Smeureanu et al., 2017; Fan et al., 2020; Sharma and Dhama, 2020] en analysant les vidéos provenant des caméras de surveillance. Les systèmes de gestion du trafic l'utilisent pour détecter les accidents [Yao et al., 2019; Nguyen et al., 2020; Khan et al., 2022]. En ce qui concerne les services de soins à domicile, la détection d'anomalies dans les vidéos aide à surveiller les personnes âgées pour assurer leur sécurité [Galvão et al., 2017; Thuc et al., 2017; Zhang et al., 2024].

Les défis principaux proviennent de leur la haute dimensionnalité et du grand volume des données vidéo, qui requiert des ressources informatiques substantielles. Les besoins de traitement en temps réel dans de nombreuses applications, comme la surveillance, ajoute une couche de complexité supplémentaire. En outre, la présence de bruit et d'artefacts, qu'ils proviennent des méthodes de compression ou des conditions environnementales, complique la tâche de distinguer les anomalies réelles des variations normales dans la vidéo.

1.5.5 Graphes

Les graphes, qui représentent les entités sous forme de nœuds et leurs relations sous forme d'arêtes, capturent des interactions complexes dans divers contextes tels que les réseaux sociaux, les réseaux de communication et les systèmes biologiques. La détection d'anomalies basée sur les graphes est particulièrement efficace dans l'analyse des réseaux sociaux pour détecter les botnets [Wang et al., 2018b; Safdari and De Bacco, 2022] et identifier les faux comptes [BalaAnand et al., 2019; Yuan et al., 2019a]. Le secteur de la cybersécurité s'appuie sur cette méthode pour la détection des intrusions [Islam et al., 2022]. Les réseaux financiers utilisent la détection d'anomalies à base de graphe pour détecter des fraudes [Li et al., 2023; Wang et al., 2023a] et des délits d'initiés [Eberle and Holder, 2009; Eberle et al., 2010], tandis qu'en biologie, il s'agit d'un processus d'analyse crucial pour comprendre les interactions entre protéines [Singh and Vig, 2017; Liu et al., 2022b].

La complexité des structures de graphes présente des défis significatifs pour la détection d'anomalies. La nature dynamique des graphes, avec des ajouts ou des sup-

pressions fréquents de nœuds et d'arêtes, exige des modèles adaptables capables de suivre le rythme des changements. De plus, l'ampleur massive des graphes, surtout dans des applications comme les réseaux sociaux et de communication, pose des défis de scalabilité. L'hétérogénéité des données de graphe, avec ses divers types et relations, complique encore le développement de techniques universelles de détection d'anomalies, nécessitant ainsi des approches adaptées pour différents scénarios.

1.6 Synthèse

Ce chapitre a exploré le domaine dynamique de la détection d'anomalies, en abordant les questions clés posées dans l'introduction de cette thèse. Nous avons d'abord clarifié la définition d'une anomalie, en nous basant sur la littérature existante, et identifié les caractéristiques principales qui les distinguent, comme leur rareté et leur divergence significative par rapport aux données normales. Le processus de détection d'anomalies a été décrit en trois étapes principales : la modélisation du comportement normal, le calcul des scores d'anomalie, et l'application d'un seuil pour identifier les anomalies.

La recherche sur la détection d'anomalies suit plusieurs paradigmes d'apprentissage, chacun ayant ses propres avantages et défis. Les méthodes non supervisées dominent traditionnellement le domaine, compte tenu de la difficulté d'obtenir des données étiquetées pour les anomalies. Cependant, les approches semi-supervisées et faiblement supervisées, qui exploitent un petit nombre d'étiquettes pour améliorer la détection, gagnent en popularité. Ces méthodes permettent de combiner la robustesse des approches non supervisées avec la précision accrue qu'offre l'utilisation de données étiquetées, même de manière limitée.

Les développements récents ont mis en lumière l'importance croissante de l'apprentissage profond dans la détection d'anomalies. Les modèles profonds, bien qu'exigeants en termes de ressources, surpassent souvent les modèles traditionnels grâce à leur capacité à gérer des données complexes et à haute dimension. En particulier, les architectures basées sur des réseaux de neurones profonds, telles que les autoencodeurs et les GANs, sont devenues des outils essentiels pour la détection d'anomalies dans des domaines comme l'imagerie et les séries temporelles.

Une autre tendance notable est l'intérêt croissant pour les techniques ensemblistes et l'apprentissage des scores de bout en bout, qui visent à améliorer la robustesse des modèles et la qualité des détections. Toutefois, ces méthodes plus sophistiquées présentent des défis en termes de ressources informatiques et d'interprétabilité, rendant le choix du modèle dépendant des contraintes spécifiques de l'application.

Ce chapitre a fourni une vue d'ensemble des concepts, méthodes et applications de la détection d'anomalies. La prochaine étape de cette thèse se concentrera sur la détection d'anomalies dans les données textuelles, approfondissant les méthodologies adaptées à la nature unique des données textuelles et explorant les défis spécifiques de ce domaine.

CHAPITRE 2

DÉTECTION D'ANOMALIES TEXTUELLES

Sommaire

2.1 Introduction

Dans le chapitre précédent, nous avons exploré le rôle crucial de la détection d'anomalies dans divers domaines, avec des applications à différents types de données. Alors que de nombreux travaux théoriques et pratiques ont été menés pour des données telles que les séries temporelles, les images, les vidéos et les graphes, la détection d'anomalies dans les textes reste un domaine relativement peu exploré. Ce chapitre est consacré à l'état de l'art concernant la détection d'anomalies textuelles, un champ situé à l'intersection de la fouille de données et du traitement automatique des langues naturelles (TALN).

Traditionnellement, la détection d'anomalies textuelles a été abordée comme une application spécifique de la détection d'anomalies dans le cadre plus large de fouille de données. Dans ce contexte, les méthodologies de fouille de données, plus précisément les techniques d'apprentissage automatique, ont joué un rôle central. Les techniques de TALN, en revanche, ont principalement assisté ces efforts en facilitant le pré-traitement des données et l'extraction de caractéristiques pour la représentation de textes.

Cependant, le développement rapide du TALN, notamment avec l'avènement des modèles de langue pré-entraînés (*Pretrained Language Models*, PLMs), a conduit à un changement de paradigme. Les techniques de TALN évoluent de leur rôle traditionnel en tant qu'assistants vers celui de solutionneurs de problèmes autonomes. L'essor des grands modèles de langue (*Large Language Models*, LLMs) renforce davantage ce changement, en positionnant les techniques de TALN, notamment les modèles de langue, comme fournisseurs potentiels de solutions générales pour les problèmes de détection d'anomalies.

Ce chapitre vise à retracer le parcours de la détection d'anomalies textuelles depuis ses débuts jusqu'à son état actuel, en mettant en évidence les contributions essentielles des avancées du TALN. Avant de plonger dans l'examen des techniques de détection d'anomalies textuelles, nous présenterons les concepts fondamentaux des anomalies textuelles et passerons en revue les ressources linguistiques existantes qui ont influencé la recherche dans ce domaine. Au vu de la nature expansive de la détection d'anomalies textuelles, qui englobe une variété de caractéristiques linguistiques et de types d'anomalies, nous conclurons ce chapitre en délimitant notre sujet de recherche. La section finale définira la portée de cette thèse, en ciblant spécifiquement le scénario de veille, afin de s'aligner sur les objectifs de cette thèse CIFRE.

À travers cet examen, nous visons à fournir un aperçu global de l'état de l'art en matière de détection d'anomalies textuelles, en mettant en lumière les avancées significatives et en identifiant les directions potentielles pour les recherches suivantes.

2.2 Anomalies textuelles

2.2.1 Anomalies textuelles - phénomènes linguistiques anormaux

Un défi fondamental dans la détection d'anomalies dans les textes consiste à définir ce qui constitue une « anomalie textuelle » et à déterminer quels phénomènes linguistiques peuvent être analysés dans le cadre de la détection d'anomalies. Une définition vague ou imprécise peut entraver cette tâche à plusieurs niveaux, conduisant à une annotation de corpus peu fiable, une détection inexacte et une évaluation inefficace.

Les anomalies sont généralement définies par leur déviation par rapport aux normes et sont caractérisées par leur rareté et leur imprévisibilité. Nous pouvons ainsi délimiter le concept d'anomalie textuelle et circonscrire le champ d'application de la détection d'anomalies textuelles selon deux perspectives : une perspective théorique, ancrée dans la définition de l'anomalie, et une perspective pratique, basée sur la nature spécifique des données anormales.

D'un point de vue théorique, la détection d'anomalies est applicable dans des **contextes où il existe des normes clairement définies ou largement acceptées**. Par exemple, en ce qui concerne l'utilisation de la langue, la norme, connue sous le nom de norme linguistique, est définie comme les usages linguistiques historiquement établis ou largement acceptés dans la pratique par les membres d'une communauté ou d'un groupe social spécifique. Cette norme englobe des règles et des patterns régissant divers éléments linguistiques tels que la phonologie, la morphologie, la syntaxe, la sémantique et la pragmatique. Ces normes peuvent être prescriptives, établies par des linguistes, ou descriptives, observées dans la pratique.

Dans ce contexte, les usages linguistiques qui s'écartent de ces normes sont considérés comme des anomalies. Ces déviations peuvent se manifester à différents niveaux, y compris lexicale, syntaxique, sémantique et pragmatique. Ainsi, la détection d'anomalies peut être employée pour des tâches telles que la détection des fautes d'orthographe, l'identification des néologismes, la correction des erreurs grammaticales, ou le diagnostic des troubles du développement du langage. Sous l'angle sociolinguistique, cette norme peut s'étendre aux éléments régissant les interactions sociales, tels que les codes linguistiques restreints, le politiquement correct et la courtoisie, ce qui rend la détection d'anomalies utile pour identifier des comportements linguistiques déviants, tels que les discours de haine. À une échelle plus large, du point de vue de la linguistique historique et évolutive, la détection d'anomalies peut également être appliquée pour détecter les changements linguistiques.

Cependant, tous les contextes concernant des données textuelles ne se prêtent pas à une approche normative. De nombreux scénarios ne permettent pas d'établir naturellement de telles normes. En pratique, compte tenu de la nature particulière des anomalies, la détection d'anomalies textuelles peut englober une plus grande variété de phénomènes linguistiques et s'appliquer aux scénarios dans lesquels :

1. **Les données positives (anomales) sont rares**, difficiles à collecter de manière exhaustive, ou insuffisants pour entraîner un modèle représentatif ; tandis que les données non positives sont abondantes et faciles à obtenir. Par exemple :
   - ☞ **Analyse des rapports médicaux.** Alors que les maladies courantes et les affections qui suivent des schémas prévisibles sont bien documentées et abondamment disponibles dans la littérature médicale, le diagnostic des maladies rares à partir de rapports médicaux reste une tâche difficile. La rareté de ces conditions, accompagnée de préoccupations éthiques et de difficultés pratiques liées à la collecte de données, rend les méthodes de diagnostic traditionnelles moins efficaces. Dans ce cas, les techniques de détection d'anomalies peuvent être employées pour identifier les écarts dans les profils de diagnostic qui suggèrent une maladie rare, renforçant ainsi le diagnostic précoce.
   - ☞ **Analyse des documents juridiques.** Dans les procédures judiciaires, les décisions des juges sont généralement en concordance avec les décisions existantes dans des cas similaires. Ces décisions cohérentes sont abondamment documentées, fournissant un cadre clair pour l'analyse juridique. En revanche, les cas où une décision de juge s'écarte significativement des précédents sont rares. Dans ce contexte, l'identification des décisions judiciaires inhabituelles ou aberrantes peut être efficacement abordée en utilisant la détection d'anomalies textuelles, mettant en lumière des pratiques contestables qui pourraient autrement passer inaperçues [Bobur et al., 2020].
2. **Les données positives (anomales) sont imprévisibles**, très variés ou en constante évolution, ce qui les rend difficiles à définir de manière claire et exhaustive. Par contre, les données non positives sont statiques ou stables, ce qui les rend faciles à définir. Par exemple :
   - ☞ **Veille stratégique.** Il s'agit de recueillir et d'analyser des informations sur les avancées technologiques, les activités des concurrents et les tendances du marché afin de soutenir la prise de décision stratégique. Les rap-

ports et les nouvelles réguliers sur les technologies existantes et les conditions habituelles du marché ont tendance à se concentrer sur un nombre limité de sujets, formant ainsi un scénario statique. Cependant, les informations les plus critiques résident généralement dans des événements inattendus ou imprévisibles, qui peuvent prendre de nombreuses formes, des dépôts de brevets aux annonces de nouveaux produits. Cette imprévisibilité et cette variabilité constituent un défi pour la définition exhaustive de toutes les possibilités potentielles. Dans de tels cas, la détection d'anomalies peut être utilisée pour identifier et évaluer ces déviations, fournissant ainsi des renseignements précieux qui pourraient autrement être négligés.

— ☞ **Sécurité de l'information.** Il s'agit de surveiller et d'analyser les journaux du système afin de détecter les failles de sécurité, les intrusions ou les activités des logiciels malveillants. Les entrées régulières dans les journaux et les activités habituelles du réseau suivent généralement des schémas et des comportements standard prédéfinis ou prévisibles. Toutefois, les intrusions peuvent varier considérablement et continuer à évoluer au fil du temps. Dans ce contexte, la détection d'anomalies textuelles est particulièrement utile. Elle permet d'identifier les écarts par rapport à ces schémas standard prédéfinis, permettant la détection rapide des anomalies dans le comportement du système qui pourraient indiquer des menaces potentielles pour la sécurité.

En résumé, une anomalie textuelle est identifiée par sa déviation par rapport aux normes établies, typiquement caractérisée par la rareté et l'imprévisibilité. Ce cadre conceptuel est généralement applicable aux contextes où les normes sont bien définies. En outre, il s'avère particulièrement pertinent dans les scénarios où les données textuelles cibles sont rares ou présentent des patterns imprévisibles.

La détection d'anomalies dans les textes trouve des applications pratiques dans divers scénarios du monde réel, tels que l'analyse de rapports financiers [Barrett et al., 2019; Lokanan et al., 2019], l'analyse des brevets [Yoon, 2012; Wang and Chen, 2019], l'analyse des rapports d'accidents [Verma and Maiti, 2018; Song and Suh, 2019], l'analyse des systèmes complexes [Srivastava and Zane-Ulman, 2005], la détection des thématiques ou tendances émergentes [Takahashi et al., 2014; Sufi, 2022], la modélisation et la détection des événements [Dasigi and Hovy, 2014; Guille and Favre, 2015; Wurzer et al., 2015; Sufi, 2022], la détection de l'extrémisme et de la radicalisation [Kramer, 2010; Agarwal and Sureka, 2015; Aldera et al., 2021], la détection des discours de haine [Gröndahl et al., 2018], la détection des fausses nouvelles [Ruchansky et al., 2017; Li et al., 2021; Garcia et al., 2023], la détection des rumeurs [Chen et al., 2018b; Torshizi and Ghazikhani, 2019], la détection des spams [Laorden et al., 2014; Wang et al., 2016a] et la détection des erreurs d'annotation [Eskin, 2000; Matoušek and Tihelka, 2017].

2.2.2 Formats de données textuelles

La détection d'anomalies textuelles peut être appliquée à des données textuelles sous différents formats, qui se divisent généralement en trois catégories : structurées, semi-structurées et non structurées. Chaque catégorie présente des défis uniques et nécessite des techniques spécifiques pour une détection efficace des anomalies.

Données structurées Les données textuelles structurées sont hautement organisées et respectent un format ou un schéma prédéfini, ce qui facilite l'analyse et le traitement. Cette catégorie englobe des sous-types tels que les données tabulaires (par exemple, les tables de bases de données, les feuilles de calcul, les fichiers CSV), les données hiérarchiques (par exemple, l'arborescence), les données de type dictionnaire (paires clé-valeur) et les données sous forme de graphes.

Une application notable de la détection d'anomalies dans les textes structurés concerne les graphes de connaissances. Dans ce contexte, les techniques de traitement du texte sont intégrées avec des méthodes de détection d'anomalies basées sur les graphes pour identifier des anomalies, telles que des relations suspectes entre entités. Une autre application concerne les données textuelles structurées en arborescence, comme les systèmes de fichiers, où la détection d'anomalies est utilisée pour identifier des éléments mal classés ou une organisation aberrante des répertoires [Fouche et al., 2020]. Cette application est cruciale pour maintenir l'intégrité organisationnelle et assurer la cohérence des données dans des systèmes de fichiers larges et complexes.

Dans ces scénarios, les informations structurelles sont généralement prioritaires par rapport au contenu textuel lui-même. L'accent est mis sur les aspects organisationnels et relationnels des données, garantissant que la structure respecte les normes attendues et que toute déviation est rapidement identifiée et corrigée. Par conséquent, l'analyse du contenu textuel occupe souvent une place marginale dans le processus de détection d'anomalies pour les données textuelles structurées.

Données semi-structurées Les données textuelles semi-structurées, bien que moins rigidement formatées que les données structurées, comportent encore des balises ou des marqueurs qui séparent les éléments sémantiques et imposent une hiérarchie des enregistrements et des champs. Ce format offre une plus grande flexibilité dans la structure tout en maintenant suffisamment d'organisation pour faciliter l'analyse et le traitement.

La détection d'anomalies dans les textes semi-structurés se concentre principalement sur les fichiers journaux. Ces fichiers sont essentiels pour la surveillance et le diagnostic des systèmes et contiennent généralement des éléments tels que des horodatages, des types de journal (par exemple, INFO, ERROR) et des messages. La structure des fichiers journaux peut varier considérablement selon les systèmes ou les événements, posant des défis uniques pour la détection d'anomalies. Le processus typique de détection d'anomalies dans les données de journal consiste à transformer le texte en une série de templates prédéfinis à l'aide de techniques de TALN, notamment l'analyse syntaxique (*parsing*) de texte. Ensuite, des algorithmes de détection d'anomalies sont appliqués pour identifier les déviations par rapport aux schémas attendus. Dans ce cas, le TALN, bien que jouant un rôle secondaire, est essentiel pour gérer la complexité et la variabilité des données de journal.

Données non structurées Les données textuelles non structurées sont le format le plus flexible mais aussi le plus complexe, car elles ne suivent aucun schéma ou structure organisationnelle prédéfinis. Ce type de données domine la recherche et les applications en TALN en raison de son omniprésence et de la richesse des informations qu'il contient. Cette catégorie englobe la grande majorité des données textuelles

générées par les humains, telles que les courriels, les posts sur les réseaux sociaux, les articles, les rapports et d'autres communications écrites.

La détection d'anomalies dans les textes non structurés nécessite souvent des techniques de TALN plus sophistiquées pour imposer une structure, extraire des caractéristiques ou analyser les contenus en vue de l'identification des anomalies. Par exemple, dans les avis de clients, les anomalies peuvent se manifester sous la forme de sentiments marginaux qui dévient considérablement de la majorité des avis ou sous la forme de contenus inauthentiques ou indésirables. De même, dans les communications par courriel, l'accent peut être mis sur l'identification des tentatives d'hameçonnage (*phishing*) ou des changements inhabituels de ton ou de style, qui pourraient indiquer des failles de sécurité ou des fraudes. Ces applications nécessitent une compréhension nuancée de la langue naturelle et des indices contextuels, ce qui place le TALN et l'analyse du contenu textuel au centre des efforts de détection d'anomalies dans les environnements de données non structurées.

Cette progression des données structurées vers les données non structurées en passant par les données semi-structurées illustre une dépendance croissante et une sophistication accrue des techniques de TALN.

2.2.3 Caractéristiques des anomalies textuelles

La détection d'anomalies dans les données textuelles présente des défis uniques en raison des propriétés inhérentes de la langue et de la communication. Une compréhension claire de ces caractéristiques est essentielle pour développer des techniques efficaces d'identification et d'interprétation des anomalies dans les textes.

Dépendance contextuelle La détection d'anomalies dans les textes est fortement dépendante du contexte. Un texte considéré comme anormal dans un scénario ou un corpus peut être tout à fait normal dans un autre. Par exemple, les articles avec des discussions détaillées sur des avancées technologiques novatrices sont attendus dans les rubriques scientifiques et technologiques, mais seraient inhabituelles dans les sections économiques ou financières. Lorsque de telles discussions apparaissent dans un contexte économique, elles peuvent suggérer des impacts de technologies émergentes sur la dynamique du marché ou indiquer des innovations disruptives.

Subjectivité La perception des anomalies textuelles est également subjective, influencée par les perspectives individuelles ou culturelles, les expériences et les biais. Par exemple, le terme « *pants* » désigne des pantalons en anglais américain, tandis qu'en anglais britannique, il fait généralement référence à des sous-vêtements. Un américain pourrait considérer comme normal d'utiliser le terme « *pants* » pour désigner des pantalons, tandis qu'un britannique pourrait trouver cet usage déroutant ou amusant. Cela illustre le défi de l'établissement de normes universelles pour la détection d'anomalies textuelles, car les différences culturelles et linguistiques doivent être prises en compte pour éviter les erreurs de classification.

Nature dynamique Dans les données textuelles, le statut d'anormalité ou de nouveauté n'est pas statique. Leur pertinence et leur interprétation peuvent changer au fil du temps, rendant commun ce qui était autrefois anormal, et vice versa. Par exemple, les premiers rapports sur le COVID-19 fin 2019 étaient initialement perçus

comme des anomalies isolées. Cependant, en mars 2020, la situation s'est transformée en pandémie mondiale, dominant les médias et les discussions publiques. Cette transition d'une anomalie à une préoccupation généralisée illustre la rapidité avec laquelle les signaux précoces peuvent évoluer en événements significatifs. Les systèmes de détection d'anomalies jouent un rôle crucial dans l'identification de ces premiers signaux, ce qui permet des réponses et des adaptations opportunes à l'évolution de la situation.

Multi-aspects Les anomalies textuelles peuvent se manifester sur plusieurs aspects simultanément, tels que le vocabulaire, la syntaxe, le sentiment, la thématique et le genre. Par exemple, dans les systèmes de surveillance de la réputation en ligne qui suivent les changements d'opinion publique ou les retours des clients, les données d'entrée peuvent contenir non seulement des changements de sentiment, mais aussi des variations en termes de sujet, de registre linguistique ou même de langue utilisée. Ces variations non pertinentes introduisent du bruit qui peut obscurcir les analyses et compliquer la détection et l'interprétation des anomalies. En particulier, les méthodes de fouille de données qui dépendent de la modélisation des distributions de données peuvent éprouver des difficultés à s'adapter à ces changements multifacettes et dynamiques.

Haute dimensionnalité Au-delà des caractéristiques inhérentes du texte lui-même, la représentation compréhensible par machine des données textuelles pose également des problèmes en raison de leur dimensionnalité typiquement élevée. Qu'elles soient transformées par des techniques telles que les sacs de mots ou les plongements de textes les plus récents, les espaces de caractéristiques des données textuelles sont généralement caractérisés par une haute dimensionnalité, allant de centaines à des milliers de dimensions. Cette complexité pose des défis importants pour les algorithmes de détection d'anomalies, en particulier ceux qui sont soumis à la « malédiction de la dimensionnalité ».

2.3 Ressources linguistiques

Dans la détection d'anomalies textuelles, les caractéristiques distinctives des anomalies textuelles, telles que la dépendance contextuelle et la subjectivité, posent non seulement des défis aux méthodologies de détection mais compliquent également le développement des ressources de support nécessaires. Ces caractéristiques contribuent à une pénurie notable, voire une absence complète, de corpus dédiés. À notre connaissance, il n'existe actuellement aucun corpus accessible au public spécifiquement conçu et annoté à cette fin.

Une solution viable consiste à utiliser directement des corpus créés pour des tâches similaires, notamment la détection de nouveautés et la détection de première occurrence d'événements. Pour mieux pallier l'absence de corpus annotés dédiés, les chercheurs ont proposé trois stratégies principales dans la littérature : la synthèse, la fusion et l'adaptation. La synthèse consiste à générer des ensembles de données artificielles basées sur des distributions et des vocabulaires spécifiques, généralement avec un accent sur la création d'instances d'anomalies. La fusion consiste à combiner des textes provenant de diverses sources, en sélectionnant des exemples « normaux » dans l'un et des anomalies dans l'autre, pour construire un ensemble de données

complet. La stratégie la plus courante, l'adaptation, consiste à modifier des corpus existants initialement conçus pour d'autres tâches afin de les rendre adaptés à la détection d'anomalies, souvent en sous-échantillonnant certaines classes pour mettre en évidence les anomalies.

Cette section examine ces stratégies en explorant quatre types de corpus : annotés, synthétisés, fusionnés et adaptés, et en explorant comment chacun contribue à surmonter les défis dans le développement des ressources pour la détection d'anomalies.

2.3.1 Corpus annotés

Les corpus annotés pour la détection de nouveautés et la détection de première occurrence d'événements (*First Story Detection*, FSD) sont étroitement liés à la détection d'anomalies textuelles. Ils fournissent des ressources précieuses pour identifier des patterns inhabituels ou des déviations dans les données textuelles, un aspect central de la détection d'anomalies.

La détection de nouveautés est étroitement liée à la détection d'anomalies, car elle se concentre sur l'identification d'informations nouvelles ou uniques au sein d'un ensemble de données. Ce processus implique intrinsèquement de repérer des anomalies, autrement dit de nouvelles données qui se distinguent des patterns connus. Un exemple typique des corpus de détection de nouveautés est le TREC Novelty Detection Corpus [Harman et al., 2002; Soboroff and Harman, 2003; Soboroff, 2004; Soboroff and Harman, 2005], conçu pour détecter les nouveautés sémantiques et syntaxiques dans les articles de presse segmentés et annotés au niveau de la phrase. L'ensemble de données plus récent, TAP-DLND 1.0 [Ghosal et al., 2018], catégorise les documents en `novel` et `non-novel`, en se concentrant sur l'élimination des informations et contenus connus pour mettre en lumière les informations nouvelles. Son successeur, TAP-DLND 2.0 [Ghosal et al., 2022], offre une analyse plus fine avec des annotations au niveau des phrases, fournissant une évaluation détaillée de la nouveauté au sein des documents.

La FSD [Allan et al., 2000], issue du projet *Topic Detection and Tracking* (TDT) [Wayne, 1997], consiste à détecter la première occurrence de nouveaux événements ou sujets, qui peuvent être considérés comme des anomalies dans le flux continu de données textuelles. Les corpus TDT, qui ont débuté avec le TDT Pilot Study Corpus [Allan et al., 1998] et se sont étendus à travers TDT2 [Cieri et al., 1999a; Fiscus et al., 1999], TDT3 [Graff et al., 1999], TDT4, et TDT5, couvrent une large éventail d'événements annotés provenant à la fois de sources d'information de la presse et de la radiodiffusion. Ces ensembles de données fournissent une riche source pour entraîner des systèmes à reconnaître des déviations significatives par rapport aux sujets ou événements connus, à l'appui des scénarios tels que la détection d'événements, l'identification de désinformation et les systèmes d'alerte précoce.

2.3.2 Corpus synthétisés

Les données synthétiques ont été largement utilisées en détection d'anomalies [Lim et al., 2018; Gu et al., 2019; Pang et al., 2019], en particulier pour tester la robustesse et l'efficacité des systèmes de détection. Au vu des complexités spécifiques associées aux données textuelles, l'application de données synthétiques dans ce contexte

a été relativement limitée, avec, toutefois, certaines tentatives en ce sens dans la littérature.

Une approche innovante est détaillée dans l'étude de Christophe et al. [2020], où un système de simulation génère des ensembles de données textuelles avec des occurrences de nouveautés contrôlées. Il s'agit de simuler des nouveautés émergentes, cycliques et ponctuelles en utilisant des modèles thématiques pour créer des textes (sacs de mots) dotés d'une dynamique temporelle spécifique. En contrôlant la fréquence et la nature de ces nouveautés, les chercheurs peuvent évaluer divers méthodes de détection d'anomalies, telles que celles basées sur la modélisation thématique ou l'analyse statistique, dans différentes conditions.

Pantin et al. [2022] a également proposé un modèle nommé GenTO qui génère des données textuelles synthétiques pour simuler des anomalies, facilitant ainsi l'évaluation des méthodes de détection d'anomalies. Il fonctionne en créant deux types d'aberrations : indépendantes et contextuelles. Les aberrations indépendantes sont générées en sélectionnant des thématiques non liés aux données normales, produisant des anomalies avec un contenu entièrement distinct. Les aberrations contextuelles, en revanche, consistent à générer des textes partageant un thème général avec les données normales mais contenant des éléments inhabituels, tels que des choix de mots rares ou des structures narratives inattendues, ce qui les rend subtilement anormaux. Cette approche permet de tester la capacité des modèles de détection à identifier à la fois des anomalies nettes et nuancées dans les ensembles de données textuelles.

En outre, Jafari [2022] a exploré l'utilisation de données synthétiques en injectant des anomalies dans un ensemble de données textuelles standard. Ce processus implique la modification d'échantillons normaux pour introduire des éléments aberrants. Cette méthode met les systèmes de détection au défi d'identifier des déviations subtiles par rapport à la norme, améliorant ainsi leur capacité à traiter une variété de scénarios anormaux.

2.3.3 Corpus fusionnés

La méthodologie de fusion, bien que peu utilisée, est également un moyen efficace de créer des corpus pour la détection d'anomalies textuelles. Cette technique consiste à combiner des textes provenant de diverses sources, telles que différents sites web, journaux ou catégories au sein d'une bibliothèque en ligne, pour compiler un ensemble de données comprenant à la fois des exemples normaux et anormaux.

Un exemple notable de cette approche est présenté dans le travail de Dasigi and Hovy [2014], qui implique la construction d'un corpus en fusionnant différentes sources de titres de journaux. Les chercheurs ont utilisé le corpus Gigaword Events, une collection de textes d'agences de presse standard, pour fournir les exemples normaux. Pour les exemples anormaux, ils ont sélectionné du contenu provenant de la rubrique « *Weird News* » du site NBC News, qui présente des événements inhabituels ou rares. Cette combinaison unique a permis aux chercheurs de créer un corpus où les données anormales déviaient significativement de la norme, facilitant ainsi l'entraînement et l'évaluation de modèles spécialement conçus pour détecter des anomalies dans les textes de presse.

2.3.4 Corpus adaptés

L'adaptation est la stratégie la plus couramment employée, qui consiste à exploiter des corpus existants initialement conçus pour d'autres fins. Il s'agit souvent de modifier ces corpus pour mettre en évidence les anomalies, généralement en réduisant l'échantillonnage de certaines classes. Cette méthode permet aux chercheurs de réutiliser des ensembles de données bien établis pour répondre aux exigences particulières de la détection d'anomalies sans avoir à créer de nouvelles données à partir de zéro. Les ensembles de données fréquemment utilisés pour l'adaptation incluent Reuters, AGNews [Zhang et al., 2015], 20NewsGroups, et IMDB [Maas et al., 2011].

Par exemple, le corpus Reuters-21578, traditionnellement utilisé pour classifier des thèmes tels que `acq` (acquisitions) et `earn` (gains), est adapté en désignant ces catégories dominantes comme des classes normales. En revanche, les catégories moins fréquentes comme `interest` sont traitées comme des anomalies, créant ainsi un scénario où la rareté du thème souligne son statut d'anomalie [Barrett et al., 2019; Yap, 2020; Hu et al., 2021; Han et al., 2022; Mai et al., 2022; Das et al., 2024]. De même, le corpus AGNews, qui classe les articles de presse en quatre catégories, à savoir `world`, `sports`, `business` et `science`, est adapté en sélectionnant une catégorie comme normale et les autres comme anormales [Manolache et al., 2021; Madan et al., 2021; Mai et al., 2022; Zeng et al., 2022; Bejan et al., 2023; Breidenstein and Labeau, 2024; Das et al., 2024]. L'adaptation du corpus 20NewsGroups utilise généralement sa hiérarchie de catégories de premier niveau. Les chercheurs peuvent choisir une large catégorie, comme `computer`, comme classe normale, tandis que les textes des autres catégories, telles que `science` ou `politics`, sont utilisés comme anomalies [Barrett et al., 2019; Hu et al., 2021; Mai et al., 2022; Bejan et al., 2023; Breidenstein and Labeau, 2024; Das et al., 2024].

Le jeu de données IMDB, initialement destiné à la classification de sentiments, offre un autre exemple d'adaptation. Ici, les commentaires étiquetés sous un sentiment (par exemple, positif) sont considérés comme normaux, tandis que ceux sous le sentiment opposé (négatif) sont considérés comme des anomalies [Ruff et al., 2019; de la Torre-Abaitua et al., 2021; Madan et al., 2021; You et al., 2021; Mai et al., 2022]. Cette approche est également appliquée à d'autres corpus d'analyse de sentiments comme Amazon [Keung et al., 2020] et Yelp, où la polarité des sentiments aide à définir des sous-ensembles normaux et anormaux.

Les progrès récents dans les méthodologies d'adaptation sont notamment marqués par la création de benchmarks comme le jeu de données AD-NLP [Bejan et al., 2023]. AD-NLP est un benchmark complet conçu pour faciliter la recherche en détection d'anomalies textuelles à travers divers types d'anomalies, y compris syntaxiques, sémantiques, pragmatiques et stylistiques. En plus d'intégrer les adaptations existantes comme AGNews et 20NewsGroups, il propose également de nouvelles adaptations comme CoLA, VUA, Song Genres et Gutenberg Categories, afin de traiter différents types d'anomalies telles que le changement de genre et la fausse paternité. Cette approche permet une évaluation complète des systèmes de détection d'anomalies, en les équipant pour identifier et analyser efficacement les déviations à travers diverses dimensions et contextes textuels.

2.4 Approches

La détection d'anomalies textuelles a traditionnellement été une application de niche dans le domaine plus large de la fouille de données. Historiquement, les méthodologies de fouille de données, en particulier les techniques d'apprentissage automatique, ont constitué la pierre angulaire de ce domaine, avec les techniques de TALN supportant ces efforts principalement par le biais du prétraitement des données et de l'extraction de caractéristiques. Au cours des dernières décennies, cependant, le paysage de la détection d'anomalies textuelles a connu une évolution significative, sous l'impulsion des progrès réalisés à la fois dans le domaine de l'apprentissage automatique et de la modélisation de la langue.

Cette progression reflète les tendances générales de l'intelligence artificielle, notamment la montée en puissance des modèles de langue préentraînés (*Pretrained Language Models*, PLMs). Ces modèles ont révolutionné le domaine en offrant des solutions sophistiquées pour la détection d'anomalies dans les données textuelles. Nous présentons ici un aperçu général des grandes phases et des techniques clés qui ont façonné le développement des méthodologies de détection d'anomalies textuelles, en mettant l'accent sur les contributions révolutionnaires des techniques de TALN.

2.4.1 Approches à base de fouille de données

Depuis les débuts de la détection d'anomalies textuelles, les méthodologies ont principalement reposé sur des techniques de fouille de données. Cette approche se déroule généralement en deux phases distinctes : d'abord, l'utilisation de techniques de TALN pour transformer le texte non structuré en une représentation compréhensible par machine ou pour extraire des caractéristiques à partir de texte brut ; ensuite, l'application à cette représentation textuelle d'algorithmes de détection d'anomalies issus du domaine la fouille de données.

Représentation de texte traditionnelle La représentation de texte traditionnelle englobe une variété de techniques, allant des méthodes simples telles que le sac de mots (*Bag of Words*, BoW), les n-grams, et la TF-IDF (*Term Frequency-Inverse Document Frequency*), à des approches plus complexes telles que les méthodes de modélisation thématique comme l'allocation de Dirichlet latente (*Latent Dirichlet Allocation*, LDA) et les méthodes de compression de données comme l'analyse en composantes principales (*Principal Component Analysis*, PCA).

Par exemple, Barrett et al. [2019] ont exploré la détection d'anomalies textuelles dans les rapports financiers, en se concentrant sur les sections liées aux facteurs de risque dans les rapports annuels des entreprises. Ils ont utilisé diverses techniques de représentation textuelle, y compris BoW, TF-IDF, et PCA pour transformer le texte brut. Ces représentations ont facilité l'utilisation d'algorithmes de fouille de données tels que KNN, LOF et OCSVM pour détecter des anomalies dans les textes.

De même, Song and Suh [2019] ont utilisé TF-IDF pour extraire des mots-clés significatifs des textes narratifs rapports d'accidents industriels. Ces mots-clés ont ensuite été analysés à l'aide de l'algorithme LOF pour identifier les anomalies basées sur les déviations de densité locale. Leur approche a permis d'identifier efficacement les accidents critiques et peu fréquents en comparant les schémas de mots-clés des incidents anormaux à ceux des incidents normaux.

Mei et al. [2018] ont exploré l'utilisation du clustering sémantique et des auto-encodeurs pour détecter des nouveautés (idées innovantes) dans les corpus de textes courts, en particulier des échanges de *brainstorming*. Ils ont utilisé la modélisation thématique, notamment LDA, pour plonger les textes dans un espace sémantique et ont ensuite utilisé des auto-encodeurs pour reconstruire ces plongements (*embeddings*). L'erreur de reconstruction a servi d'indicateur d'anomalie, des erreurs plus élevées suggérant des idées plus novatrices ou anormales.

Plongements statiques Avec les progrès des techniques de TALN dans les années 2010, la détection d'anomalies textuelles a vu l'intégration de méthodes de représentation de texte plus sophistiquées, allant des plongements de mots statiques aux plongements de documents contextuels. Durant cette période, bien que les algorithmes de fouille de données demeurent centraux, les techniques de TALN ont commencé à jouer un rôle plus important dans l'amélioration des performances globales.

L'introduction de plongements de mots statiques pré-entraînés tels que Word2Vec Mikolov et al. [2013], GloVe [Pennington et al., 2014] et FastText [Bojanowski et al., 2017] a considérablement fait progresser les efforts de détection d'anomalies textuelles. Ces modèles ont permis une représentation plus nuancée des données textuelles en capturant les associations de mots basées sur leur utilisation dans de vastes corpus.

Seo et al. [2020] ont développé un système pour identifier les réponses inhabituelles des clients dans les enquêtes sur la fiabilité des véhicules. Ils ont utilisé Doc2Vec [Le and Mikolov, 2014], une technique de plongements de texte basée sur un réseau de neurones, pour convertir les commentaires des clients en vecteurs continus. L'algorithme LOF a ensuite été appliqué à ces vecteurs pour détecter les réponses anormales. De plus, l'étude a utilisé TF-IDF pour extraire des mots-clés significatifs des textes anormaux, aidant à visualiser et interpréter les préoccupations inhabituelles des clients, permettant ainsi aux ingénieurs de traiter les retours critiques.

Cichosz [2020] a exploré la détection non supervisée d'anomalies dans les forums en ligne, en utilisant des plongements de mots pour la représentation des textes. L'auteur a utilisé les plongements GloVe (*Global Vectors*) pour convertir les messages des forums en représentations vectorielles denses. Pour la détection des anomalies, l'étude a appliqué deux méthodes : OCSVM et k-medoids clustering. Les résultats ont démontré que les plongements GloVe, combinés à l'approche de dissimilarité des clusters, surpassaient les modèles BoW traditionnels, offrant une meilleure qualité de détection et atténuant les problèmes de haute dimensionnalité souvent rencontrés dans le clustering de texte.

Ait-Saada and Nadif [2023] se sont concentrés sur la détection d'anomalies au niveau de thématiques dans les textes courts. Ils ont utilisé une combinaison de plongements FastText, qui fournissent des représentations vectorielles denses pour les mots, et de modèles de mélange gaussien (GMM) pour capturer les variations sémantiques au sein des textes. En exploitant la capacité des modèles GMM à modéliser la distribution de probabilité des plongements, l'étude a identifié efficacement les échantillons anormaux qui déviaient significativement de la distribution normale du corpus.

Plongements contextuels L'évolution s'est poursuivie avec l'incorporation d'architectures d'apprentissage profond, conduisant au développement de plongements dynamiques et contextuels comme ELMo (*Embeddings from Language Models*) et BERT (*Bidirectional Encoder Representations from Transformers*). Ces modèles fournissent des représentations de mots et de phrases plus riches et conscientes du contexte, en capturant des nuances sémantiques que les méthodes précédentes ne pouvaient pas saisir. Cette progression dans la représentation de texte a révolutionné davantage la détection d'anomalies textuelles, ce qui permet une identification plus précise des anomalies en s'appuyant sur une compréhension plus profonde de la sémantique et du contexte fournie par ces modèles.

Madan et al. [2021] ont proposé le cadre TADPOLE pour aborder le problème du changement de domaine souvent rencontré dans le paradigme de « pré-entraînement et ajustement ». Ce cadre vise à réaliser un pré-entraînement adapté à la tâche cible grâce à des techniques de détection d'anomalies. Pour y arriver, ils ont utilisé des plongements BERT pour représenter les textes et ont employé des algorithmes tels que PCA, LOF et Isolation Forests pour détecter des anomalies. Il s'agit d'identifier les données spécifiques au domaine cible à partir d'un corpus général en classant les textes en fonction de leurs scores de pertinence dérivés de la détection d'anomalies. Cette méthode permet une adaptation efficace au domaine, particulièrement utile pour les tâches avec des données étiquetées limitées. En filtrant les informations moins pertinentes, TADPOLE améliore efficacement la performance des modèles de langue dans les tâches en aval.

Kumar et al. [2022] ont développé un modèle de détection d'anomalies pour les données de Twitter, en se concentrant sur les tweets liés à la santé pendant la pandémie de COVID-19. Ils ont utilisé une combinaison de LDA et de NMF (*Non-negative Matrix Factorization*) pour la modélisation thématique, créant ainsi un ensemble de requêtes concernant les thématiques dominantes. Par la suite, le modèle BERT a été appliqué pour évaluer la similarité sémantique des nouveaux tweets par rapport à ces requêtes. Le clustering K-means a été utilisé pour regrouper des tweets similaires, distinguant efficacement les schémas de tweets habituels et inhabituels, identifiant ainsi les anomalies dans les discussions sur la santé publique.

Cadre intégré Plus récemment, le passage des étapes séparées de représentation de texte et de détection d'anomalies à un cadre plus unifié a constitué une évolution notable. Un exemple typique est le modèle CVDD (*Context Vector Data Description*) introduit par Ruff et al. [2019]. CVDD utilise des modèles de plongements de mots pré-entraînés, tels que GloVe ou FastText, pour transformer les mots en représentations vectorielles de longueur fixe, capturant des informations sémantiques générales à partir de grands corpus. Le mécanisme d'auto-attention multi-têtes opère ensuite sur ces plongements, pour créer des combinaisons pondérées qui aboutissent à de multiples plongements de phrases spécifiques au contexte. Chaque tête d'attention se concentre sur différentes parties de la phrase, ce qui permet d'extraire divers aspects sémantiques. Ces plongements spécifiques au contexte sont comparés à un ensemble de vecteurs de contexte appris, que le modèle affine pendant l'entraînement. L'objectif est de minimiser la distance cosinus entre les plongements de phrases et ces vecteurs de contexte, regroupant efficacement les données normales autour de ces représentations apprises. Les points de données anormaux sont identifiés comme ceux dont les plongements ne s'alignent pas bien avec les vecteurs de contexte, indiquant

qu'ils dévient des patterns normaux. Ce modèle intégré permet à CVDD de réaliser une détection d'anomalies contextuelles en exploitant les informations sémantiques riches et pré-entraînées et en les affinant pour capturer les différences nuancées, ce qui améliore à la fois la fiabilité et l'interprétabilité de la détection.

De même, le cadre FATE (*Few-shot Anomaly Ddetection in TExt with deviation learning*), introduit par Das et al. [2024], avance davantage cette approche unifiée. Contrairement à CVDD, FATE utilise un modèle basé sur les transformeurs, spécifiquement BERT, pour générer des plongements dynamiques et contextuels des textes d'entrée. Ces plongements sont traités par une couche d'auto-attention multi-têtes pour produire plusieurs scores d'anomalie capturant différents aspects sémantiques du texte. FATE apprend explicitement les scores d'anomalie via une approche d'apprentissage par déviation. Les échantillons normaux sont guidés pour s'aligner étroitement avec un score de référence dérivé d'une distribution préalable, tandis que les échantillons anormaux sont conçus pour dévier significativement de cette référence. Cette approche permet à FATE d'utiliser efficacement un nombre limité d'anomalies étiquetées, offrant un cadre robuste et évolutif pour la détection d'anomalies dans divers ensembles de données textuelles. En intégrant la génération de plongements et le calcul de scores d'anomalies dans un seul modèle, FATE améliore à la fois la précision et l'interprétabilité des résultats de détection d'anomalies.

Extraction de caractéristiques À part la représentation de texte, les techniques de TALN servent également à l'extraction de caractéristiques dans des applications de détection d'anomalies plus complexes, telles que la détection d'événements et la détection de fausses nouvelles. Dans ces scénarios, le TALN est utilisé pour extraire des informations telles que la polarité des sentiments, les mots-clés, les entités nommées et les relations. Par exemple, AI-GlobalEvents, développé par Sufi [2022], utilise des techniques de TALN comme la reconnaissance des entités nommées et l'analyse des sentiments pour extraire des caractéristiques cruciales dans de vastes quantités de textes d'actualité agrégées à partir de sources en ligne. Il s'agit d'identifier les entités clés impliquées dans les événements, de déterminer leurs relations et d'évaluer le sentiment associé à ces entités. Les caractéristiques extraites sont ensuite traitées à l'aide d'algorithmes de détection d'anomalies, en particulier le résidu spectral (*Spectral Residual*, SR) avec les réseaux de neurones convolutifs (*Convolutional Neural Networks*, CNN). Ces algorithmes analysent les données pour identifier des schémas inhabituels ou des anomalies dans le flux d'actualités, ce qui peut indiquer des événements importants ou inattendus.

Pour les approches de détection d'anomalies textuelles basées sur la fouille de données, les techniques de TALN jouent principalement un rôle de support. Leur objectif principal est d'améliorer la performance des algorithmes de détection d'anomalies en fournissant des entrées plus claires, plus significatives et mieux structurées.

2.4.2 Approches à base de modèles de langue

Modèles de langue pré-entraînés Depuis 2018, l'essor des modèles de langue pré-entraînés (*Pretrained Language Models*, PLMs) tels que BERT [Devlin et al., 2019], RoBERTa [Liu et al., 2019b] et ELECTRA [Clark et al., 2020] a marqué un changement de paradigme significatif dans le domaine du TALN, passant du paradigme traditionnel d'« entraînement et test » au paradigme de « pré-entraînement et ajustement (*pre-training and fine-tuning*) » [Liu et al., 2021b]. Ce nouveau paradigme

permet aux modèles d'exploiter les riches connaissances contextuelles encodées lors du pré-entraînement et de les adapter à des tâches en aval par le biais d'un ajustement raffiné, ce qui conduit à des avancées significatives dans diverses tâches de TALN. Ce changement de paradigme a également profondément impacté la détection d'anomalies textuelles, avec des techniques de TALN, plus précisément les modèles de langue, évoluant d'une simple assistance à la résolution de problèmes à des fournisseurs de solutions autonomes. Ce changement a été notamment exemplifié par la méthode DATE proposée par Manolache et al. [2021].

DATE (*Detecting Anomalies in Text using ELECTRA*) exploite les capacités des modèles de langue pré-entraînés dans un contexte auto-supervisé pour la détection d'anomalies textuelles. Il utilise le modèle ELECTRA [Clark et al., 2020] et se concentre sur deux tâches : la détection de tokens remplacés (*Replaced Token Detection*, RTD) et la détection de masques remplacés (*Replaced Mask Detection*, RMD). Dans la RTD, un générateur remplace certains tokens par des alternatives plausibles, et le discriminateur du modèle prédit si chaque token est original ou remplacé. Cette tâche aide le modèle à comprendre les distributions typiques des tokens. La RMD étend le concept aux séquences, où le modèle prédit le schéma de masque utilisé lors du remplacement des tokens. Il s'agit de masquer certains mots dans une phrase et de les remplacer, puis de demander au discriminateur de déterminer le schéma de masque original. La combinaison de ces tâches permet à DATE de capturer efficacement les normes au niveau des tokens et des séquences. Le modèle génère des scores d'anomalie basés sur un score de pseudo-étiquette, calculé en agrégant la confiance des résultats du discriminateur. Ce score indique à quel point un texte dévie des schémas normaux appris, des scores plus élevés suggérant des anomalies potentielles. Ce cadre illustre le changement de paradigme où des modèles de langue comme ELECTRA non seulement assistent à l'extraction de caractéristiques, mais jouent également un rôle central dans la détection d'anomalies, agissant ainsi comme des fournisseurs de solutions autonomes dans le processus de détection d'anomalies.

Le modèle DATE est devenu un point de référence important dans la recherche sur la détection d'anomalies textuelles [Han et al., 2022; Ait-Saada and Nadif, 2023; Das et al., 2024]. Cependant, une limitation significative de cette méthode est son accent sur les applications académiques plutôt que sur le déploiement pratique et réel. Le cadre original de DATE ne fournit pas de prédictions directes sous forme de scores d'anomalie ou d'étiquettes binaires. Au lieu de cela, il génère des sorties sous forme de mesures d'évaluation telles que le score AUCROC, qui est précieux pour la recherche et les études comparatives mais moins applicable dans des contextes opérationnels où des prédictions exploitables sont nécessaires.

Un autre exemple de l'utilisation des PLMs dans la détection d'anomalies textuelles est fourni par Mai et al. [2022]. Leur étude examine l'apprentissage auto-supervisé pour la détection d'anomalies textuelles à classe unique, en se concentrant sur des scénarios où seules des données normales sont disponibles lors de l'entraînement. Les chercheurs ont ajusté des modèles de transformeur pré-entraînés, spécifiquement BERT, en utilisant divers objectifs d'apprentissage auto-supervisé tels que la modélisation de langue masquée, la modélisation de langue causale et l'apprentissage contrastif. Ils ont utilisé les valeurs de perte issues de ces processus d'ajustement comme scores d'anomalie, sous l'hypothèse que les échantillons normaux produiraient des scores de perte inférieurs par rapport aux anomalies, qui ne se conformeraient pas à la représentation apprise par le modèle. Les résultats montrent que

cette approche d'ajustement identifie efficacement les anomalies, en particulier dans la détection des déviations sémantiques et syntaxiques. Cette recherche met en évidence comment l'utilisation des PLMs pour la détection d'anomalies non seulement améliore l'exactitude mais également simplifie le processus de détection.

Grands modèles de langue La dernière évolution en TALN a été marquée par l'avènement des grands modèles de langue (*Large Language Models*, LLMs) comme GPT-4 et LLaMA 3. Ces modèles non seulement continuent la tendance du paradigme « pré-entraînement et ajustement », mais introduisent également un nouveau paradigme, à savoir le « pré-entraînement et prompts ». Ce paradigme exploite la grande quantité de données et de connaissances générales que ces modèles acquièrent au cours du pré-entraînement. En utilisant des prompts, des tâches spécifiques peuvent être abordées en conditionnant la réponse du modèle sur le prompt donné, guidant efficacement le modèle pour générer des sorties contextuellement pertinentes et spécifiques à la tâche sans nécessiter un ajustement complet sur de nouveaux ensembles de données.

Les LLMs ont démontré leur efficacité dans de nombreuses tâches de TALN au cours des dernières années. Cependant, l'exploration des méthodes basées sur les prompts pour la détection d'anomalies textuelles est encore à ses débuts. Les recherches existantes se concentrent principalement sur des modèles tels que BERT et GPT-2, qui sont généralement classés comme des modèles de langue pré-entraînés plutôt que comme des grands modèles de langue au sens strict. De plus, les premières explorations se sont principalement concentrées sur le traitement des logs en raison de leurs schémas bien définis et de leur nature semi-structurée [Boutalbi et al., 2023; Su et al., 2024]. En ce qui concerne les textes non structurés, en particulier les documents plus longs tels que les articles de presse et les blogs, l'application des LLMs comme GPT-4, LLaMA 3 et Claude 3 dans la détection d'anomalies reste largement inexplorée. Un défi significatif dans ce domaine est les limitations de longueur de contexte des modèles, ce qui limite leur efficacité à effectuer un apprentissage en contexte pour la détection d'anomalies textuelles.

2.5 Détection d'anomalies dans la veille

Comme indiqué au début de cette thèse, l'objectif principal de cette recherche est l'application de la détection d'anomalies textuelles dans le contexte de la veille. La veille est le processus systématique de collecte, d'analyse et d'utilisation d'informations pour anticiper et répondre aux changements dans l'environnement extérieur. Elle englobe diverses activités d'intelligence visant à surveiller les tendances, les innovations et les développements dans des domaines spécifiques d'intérêt, afin de soutenir la prise de décision stratégique en identifiant les opportunités, les risques et les avantages concurrentiels.

La veille peut être largement catégorisée en plusieurs types, chacun avec un objectif spécifique. La veille technologique implique la surveillance des avancées technologiques, fournissant des informations sur les innovations émergentes susceptibles d'avoir un impact sur les différentes industries. La veille concurrentielle se concentre sur le suivi de la dynamique des concurrents, aidant les organisations à comprendre leur position sur le marché et à anticiper les stratégies de leurs rivaux. La veille commerciale observe les tendances du marché, les préférences des consommateurs et

les indicateurs économiques, offrant une vue d'ensemble des conditions du marché et des potentiels changements dans le comportement des consommateurs. La veille stratégique, quant à elle, évalue l'environnement stratégique global, y compris les facteurs politiques, économiques, sociaux et technologiques, assurant que les organisations sont conscientes des influences externes plus larges qui pourraient affecter leur direction stratégique.

L'un des principaux objectifs des systèmes de veille est la détection des signaux faibles, c'est-à-dire des indications précoces de changements potentiels ou de tendances émergentes qui peuvent devenir significatives à l'avenir. Les signaux faibles sont souvent subtils, ambigus et pas encore complètement formés, ce qui les rend difficiles à détecter et à interpréter. Ces signaux se manifestent typiquement par des déviations mineures, des événements inhabituels ou des schémas inattendus dans les données, provenant d'une large gamme de matériaux, y compris les publications scientifiques, les réseaux sociaux, les retours des clients et des sources non conventionnelles telles que les forums et les discussions de groupes d'intérêt spécial.

Ainsi, les signaux faibles peuvent être interprétés comme des types spécifiques d'anomalies textuelles, à savoir des textes véhiculant des informations inhabituelles qui suggèrent des tendances ou des changements futurs significatifs. Le processus de détection automatique des signaux faibles peut être divisé en deux phases principales :

1. **Détection des anomalies textuelles** : Cette phase initiale consiste à identifier les déviations par rapport aux schémas textuels normaux à travers diverses sources de données.
2. **Évaluation des mégatendances potentielles** : La deuxième phase analyse ces anomalies textuelles détectées pour déterminer lesquelles pourraient potentiellement se transformer en tendances significatives ou en changements majeurs.

Cette thèse se concentrera spécifiquement sur les anomalies textuelles au niveau des thématiques, en se basant sur l'idée que ces anomalies sont la principale source de signaux faibles potentiels, en particulier dans des contextes tels que la veille stratégique, concurrentielle et technologique [Hiltunen, 2007].

En conclusion, la détection d'anomalies textuelles est essentielle pour le fonctionnement efficace des systèmes de veille, facilitant l'identification précoce et l'évaluation des signaux faibles. Cette fonctionnalité renforce la capacité d'une organisation à répondre stratégiquement aux opportunités et aux menaces potentielles, garantissant qu'elle reste proactive plutôt que réactive dans un environnement en constante évolution.

2.6 Synthèse

Ce chapitre a abordé le domaine spécifique de la détection d'anomalies dans les textes, en se concentrant sur les défis uniques que présentent les données textuelles, les méthodes développées pour relever ces défis, et le rôle significatif des modèles de langue modernes dans ce domaine en pleine évolution.

Les anomalies textuelles, contrairement aux autres types de données, présentent des défis uniques dus à leur dépendance contextuelle, leur subjectivité, et leur nature dynamique. Ces propriétés rendent la détection d'anomalies textuelles particu-

lièrement complexe, nécessitant des techniques avancées de TALN. Les approches traditionnelles basées sur la fouille de données ont évolué pour intégrer des représentations textuelles plus sophistiquées, allant des méthodes simples comme le sac de mots aux plongements contextuels avancés offerts par les modèles de langue pré-entraînés.

Les approches basées sur les modèles de langue ont marqué une avancée significative dans la détection d'anomalies textuelles, en particulier avec l'émergence des PLMs comme BERT et ELECTRA. Ces modèles utilisent la compréhension contextuelle profonde et les connaissances acquises lors du pré-entraînement pour identifier les anomalies avec une efficacité accrue, ce qui illustre un changement de paradigme important dans ce domain, où les modèles de langue passent d'un rôle de support à celui de solutionneurs autonomes de problèmes.

Par ailleurs, le développement des ressources linguistiques pour la détection d'anomalies textuelles reste un défi majeur. Ce chapitre a également examiné diverses stratégies pour pallier le manque de corpus spécifiques, notamment l'utilisation de corpus annotés, synthétisés, fusionnés, et adaptés. Ces ressources sont essentielles pour entraîner et évaluer les systèmes de détection d'anomalies, permettant d'identifier des déviations textuelles significatives dans divers contextes.

En conclusion, ce chapitre a souligné les progrès essentiels réalisés dans le domaine de la détection d'anomalies textuelles, avec un accent particulier sur le rôle des techniques de TALN. La transition en cours vers des modèles de langue toujours plus puissants constitue une direction prometteuse pour les recherches futures. Bien que les grands modèles de langue tels que GPT-4 et LlaMA 3 n'aient pas encore pleinement démontré leurs capacités dans ce domaine, ils offrent un potentiel pour de nouvelles approches de détection et d'interprétation des anomalies textuelles. Cette thèse explorera plus en profondeur ces perspectives et proposera des pistes pour valider leur efficacité dans ce domaine spécifique.

Conclusion de la première partie

Dans cette première partie de la thèse, nous avons mené une revue globale de l'état de la recherche en détection d'anomalies, en mettant particulièrement l'accent sur les données textuelles. Cette exploration a couvert divers domaines de la détection d'anomalies, aboutissant à une étude approfondie des techniques spécifiques à la détection d'anomalies textuelles, qui est au cœur de notre recherche. Grâce à cette analyse, nous avons identifié les principales méthodologies, tendances et défis qui façonnent actuellement l'état de l'art.

* * *

Le chapitre 1 a fourni un aperçu général de la détection d'anomalies à travers divers domaines, en retraçant son évolution depuis les premières méthodes statistiques jusqu'aux approches contemporaines basées sur l'apprentissage automatique. Nous avons observé trois tendances de recherche significatives : le recours croissant aux réseaux neuronaux et aux modèles profonds, le changement de paradigme vers l'apprentissage semi-supervisé et faiblement supervisé, et le développement de méthodes de calcul de scores innovantes, notamment celles qui reposent sur les cadres d'apprentissage ensembliste et d'apprentissage de bout en bout. Ces tendances démontrent la nature dynamique de la recherche en détection d'anomalies, reflétant comment les progrès en disponibilité des données et en puissance de calcul ont conduit au développement de modèles sophistiqués capables de gérer des environnements complexes.

Le chapitre 2 s'est ensuite concentré sur la détection d'anomalies dans les textes, un domaine caractérisé par les défis uniques posés par la variabilité et la richesse de la langue naturelle. Nous avons exploré la manière dont les méthodes traditionnelles de détection d'anomalies ont été adaptées aux données textuelles, principalement par l'utilisation de modèles de langue et de techniques de plongements qui facilitent une représentation efficace du texte pour la détection d'anomalies. Nous avons également discuté de l'impact significatif des modèles de langue pré-entraînés, en soulignant comment ceux-ci ont permis aux techniques TALN de devenir des solutions autonomes pour la détection d'anomalies. Ce chapitre a souligné l'indépendance croissante des techniques TALN dans la gestion des défis de détection d'anomalies, une tendance qui est devenue de plus en plus prédominante avec l'avancement des grands modèles de langage. Ce chapitre a mis en évidence l'indépendance croissante des techniques TALN pour relever les défis de la détection d'anomalies, une tendance de plus en plus marquée à mesure que les grands modèles de langue deviennent plus puissants et plus sophistiqués.

* * *

Ensemble, ces chapitres ont non seulement consolidé notre compréhension des fondements théoriques et des avancées méthodologiques en détection d'anomalies, mais ont également mis en lumière des lacunes importantes et des tendances émergentes. Ces aperçus informent directement les deux axes de recherche de notre travail : premièrement, l'adaptation de méthodologies établies dans des domaines plus larges aux données textuelles, et deuxièmement, l'utilisation de techniques TALN de pointe, en particulier les grands modèles de langue, pour mettre au point de nouvelles approches dans la détection d'anomalies textuelles.

Deuxième partie

Méthodes de fouille de données

Introduction de la deuxième partie

Cette deuxième partie de la thèse vise à faire le pont entre les méthodologies de fouille de données et les spécificités des données textuelles. Cela implique une adaptation des algorithmes de détection d'anomalies existants pour répondre aux propriétés distinctes des données textuelles, suivie d'une évaluation complète de leur performance.

* * *

Le Chapitre 3 présente la méthodologie employée pour appliquer les techniques de fouille de données à la détection d'anomalies dans les données textuelles. Ce chapitre commence par discuter des approches de prétraitement et de représentation des textes, puis explore les algorithmes de détection d'anomalies adaptés aux données textuelles. L'accent est mis sur trois aspects critiques : le paradigme d'apprentissage utilisé, l'intégration des réseaux neuronaux, et les bases théoriques qui guident le calcul des scores d'anomalie.

Le Chapitre 4 se concentre sur les ensembles de données utilisés pour nos expérimentations. Il décrit les critères de sélection et les processus d'adaptation des jeux de données textuels pour la détection d'anomalies, visant à garantir la diversité et la pertinence des types d'anomalies inclus dans les jeux de données. Ce chapitre établit une base solide pour les expériences à venir en assurant une couverture adéquate des différents types d'anomalies textuelles.

Le Chapitre 5 couvre les expériences réalisées pour évaluer l'efficacité des différentes méthodes de fouille de données appliquées aux textes. Nous y présentons une comparaison détaillée des performances des différents modèles. Les résultats obtenus fournissent des perspectives clés sur les stratégies optimales pour divers scénarios de détection d'anomalies textuelles.

* * *

Ces chapitres sont guidés par les questions de recherche suivantes, qui visent à approfondir notre compréhension des défis et des opportunités liés à l'application des techniques de fouille de données à la détection d'anomalies textuelles :

— Quel est l'impact des différents **paradigmes d'apprentissage** automatique sur la détection d'anomalies textuelles, notamment lors de l'intégration de données partiellement étiquetées ? Combien d'échantillons annotés sont nécessaires pour améliorer significativement les résultats, et quels types d'anomalies annotées sont les plus utiles pour l'entraînement des modèles ?

— Comment différentes **techniques de représentation** influencent-elles l'efficacité de la détection d'anomalies dans le texte ?

— Existe-t-il des avantages à utiliser certains types de **scores d'anomalie** par rapport à d'autres, et si oui, pour quelles raisons ?
— Les modèles d'**apprentissage profond** surpassent-ils systématiquement les modèles traditionnels peu profonds dans la détection d'anomalies textuelles, comme cela a été observé dans d'autres scénarios de l'apprentissage automatique ?
— Quel est le **meilleur compromis** entre efficacité temporelle, exactitude des prédictions et exigences matérielles pour différents scénarios de détection d'anomalies textuelles ?

En répondant à ces questions, cette partie vise à offrir une évaluation globale des méthodes existantes et à proposer de nouvelles perspectives de recherche et d'innovation dans le domaine de la détection d'anomalies textuelles.

CHAPITRE 3

MÉTHODOLOGIE

Sommaire

3.1 Introduction

Dans ce chapitre, nous présentons la méthodologie employée pour détecter les anomalies textuelles dans le cadre de fouille de données, en nous concentrant sur deux composants principaux : 1) les techniques de représentation de texte et 2) les algorithmes de détection d'anomalies

La détection d'anomalies dans les textes est un défi particulièrement complexe en raison de la nature intrinsèquement non structurée et linguistiquement nuancée des données textuelles. Contrairement aux données numériques ou bien structurées, qui peuvent être traitées plus directement pour la détection d'anomalies, les données textuelles nécessitent une transformation en un format compréhensible par la machine, qui capture les subtilités sémantiques et syntaxiques. Cette transformation est au cœur du premier composant, la représentation de texte. Nous abordons comment encoder efficacement les caractéristiques linguistiques telles que les nuances, les synonymes et les ambiguïtés, établissant ainsi les bases fondamentales pour l'analyse ultérieure.

Le deuxième composant concerne les algorithmes de détection d'anomalies. Une fois que les données textuelles sont transformées en représentations numériques appropriées, typiquement sous forme de vecteurs à haute dimension, le défi est désormais de répondre aux problèmes posés par cette haute dimensionnalité à l'aide des

algorithmes de fouille de données. Ces algorithmes apprennent à partir de l'espace de caractéristiques établi par le premier composant et calculent un score d'anomalie continu pour chaque texte, basé sur un cadre théorique spécifique. Pour faciliter les analyses ultérieures, la discussion concernant les algorithmes mettra en lumière trois aspects critiques : le paradigme d'apprentissage utilisé, l'intégration de réseaux neuronaux, et les bases théoriques qui guident le calcul des scores d'anomalie.

À travers une exploration détaillée de ces méthodologies, ce chapitre vise à fournir un aperçu complet des techniques essentiels pour détecter les anomalies dans les textes.

3.2 Représentation de texte

Dans la détection d'anomalies, les caractéristiques uniques des données textuelles posent des défis importants par rapport aux données typiques pour lesquelles la plupart des algorithmes de détection d'anomalies sont conçus, telles que les séries temporelles, les signaux, les images ou les données tabulaires. Contrairement à ces formats structurés, le texte brut est intrinsèquement non structuré, avec une longueur et une complexité qui varient considérablement. De plus, l'interprétation et la compréhension des données textuelles sont fortement influencées par le contexte. Cette complexité nécessite l'emploi de techniques spécialisées de représentation pour convertir le texte en formats numériques et structurés, que les algorithmes d'apprentissage automatique peuvent alors exploiter efficacement. Une telle représentation assure que les subtilités contextuelles et nuancées du texte sont correctement capturées, facilitant ainsi une détection exacte des anomalies.

La représentation de texte est une technique fondamentale en traitement automatique des langues naturelles (TALN) et a considérablement évolué tout au long de l'histoire du domaine. Elle a commencé avec l'**Encodage One-Hot**, où chaque mot du vocabulaire est attribué un vecteur unique dans un espace de haute dimension, principalement rempli de zéros à l'exception d'un seul « *One* » à l'indice correspondant au mot. Malgré sa nature simple, la grande dimensionnalité et l'incapacité à capturer les relations entre les mots avec l'encodage One-Hot ont nécessité le développement de méthodes plus avancées. Cela a mené au modèle **Sac de Mots** (*Bag of Words* ou BoW), qui a amélioré la représentation textuelle en comptant la fréquence d'apparition de chaque mot dans un document. Bien que BoW constitue un progrès, il néglige toujours la syntaxe et la sémantique du langage, traitant le texte comme de simples collections de mots sans égard à leur ordre ou à leurs relations contextuelles.

L'introduction des **plongements de mots** (*word embeddings*) comme Word2Vec [Mikolov et al., 2013] et GloVe [Pennington et al., 2014] a marqué une progression significative. Ces techniques apprennent des représentations denses et de faible dimension des mots basées sur leurs cooccurrences contextuelles dans de grands corpus, capturant des relations sémantiques subtiles bien plus efficacement que BoW. Cependant, ces plongements statiques ne peuvent pas rendre compte des significations variables que les mots peuvent avoir dans différents contextes. Pour résoudre cela, des **plongements dynamiques ou contextuels** comme ELMo [Peters et al., 2018] et BERT [Devlin et al., 2019] ont été développés. Ces modèles peuvent générer différents plongements pour un mot en fonction de son contexte environnant. Cette avancée a considérablement amélioré la gestion des mots ayant plusieurs significations et a boosté la performance à travers diverses tâches de TAL.

Les avancées récentes ont conduit à des techniques telles que Doc2Vec [Le and Mikolov, 2014] et Sentence-BERT [Reimers and Gurevych, 2019] pour représenter des unités textuelles plus grandes telles que des phrases et des documents. Ces **plongements de phrases/documents** encapsulent des informations contextuelles plus larges, cruciales pour des tâches nécessitant la compréhension de textes entiers, comme la classification de documents.

Dans notre recherche, nous examinerons l'impact des techniques de représentation sur la détection d'anomalies textuelles en comparant des méthodes traditionnelles à des méthodes plus récentes. Nous utilisons TF-IDF, une méthode largement utilisée tant dans le milieu académique qu'industriel, comme technique traditionnelle. Pour une méthode plus récente, nous employons Sentence-BERT, qui a démontré son efficacité dans de nombreuses tâches de TALN.

3.2.1 TF-IDF

TF-IDF (*Term Frequency–Inverse Document Frequency*) est une mesure statistique utilisée pour évaluer l'importance d'un terme contenu dans un document, relativement à une collection ou un corpus. Cette méthode est largement utilisée dans la recherche d'informations et la fouille de textes. Le modèle BoW, prédécesseur du TF-IDF, compte simplement les occurrences de chaque terme dans un document pour représenter son contenu, ce qui présente un inconvénient majeur : elle traite chaque terme comme étant également important et ne tient pas compte de la fréquence des termes à travers le corpus. Cela conduit à ce que les mots courants (par exemple, "le", "est") submergent la représentation du document, obscurcissant l'importance des mots plus significatifs et spécifiques au sujet. La motivation derrière le TF-IDF est de pallier ce problème en introduisant une mesure de la spécificité d'un terme pour un document au sein d'un corpus plus large. L'objectif est de prioriser les termes qui sont fréquents dans un document particulier mais peu utilisés dans les autres documents, mettant ainsi en évidence les sujets ou termes uniques de chaque document.

Le TF-IDF est le produit de deux statistiques :

Fréquence du Terme (TF) La TF mesure la fréquence à laquelle un terme se produit dans un document. Elle peut être calculée de différentes manières, mais la plus simple consiste à compter le nombre d'occurrences d'un terme t dans un document d et à le diviser par le nombre total de termes dans le document.

$$TF(t,d) = \frac{\text{Nombre d'occurrences de terme } t \text{ dans le document } d}{\text{Nombre total de termes dans le document } d}$$

Fréquence Inverse du Document (IDF) L'IDF évalue l'inverse de la fréquence des termes à travers un corpus, ce qui aide à diminuer la significativité des termes qui apparaissent plus fréquemment dans les documents.

$$IDF(t,D) = \log\left(\frac{\text{Taille de la collection de documents } D}{\text{Nombre de documents contenant le terme } t + 1}\right)$$

Le "+1" au dénominateur évite la division par zéro et lisse les cas où un terme pourrait ne pas apparaître dans aucun document.

Fréquence du Terme - Fréquence Inverse du Document Le score TF-IDF est obtenu en multipliant ces deux statistiques :

$$TF - IDF(t, d, D) = TF(t, d) \times IDF(t, D)$$

Le TF-IDF offre des avantages significatifs tels que sa capacité à prioriser la pertinence plutôt que la fréquence pure, ce qui aide à distinguer les mots importants dans les documents. Cette caractéristique est particulièrement bénéfique pour des tâches telles que la classification de documents et le clustering, car le TF-IDF aide à sélectionner des caractéristiques descriptives qui encapsulent l'essence des documents. De plus, son efficacité computationnelle et sa scalabilité le rendent idéal pour gérer de grands jeux de données. Cependant, l'approche n'est pas sans défauts. Les calculs de TF-IDF ignorent souvent les relations contextuelles entre les mots, traitant les termes comme des unités indépendantes sans considérer les synonymes ou les connexions sémantiques, ce qui peut conduire à une perte de sens nuancé. De plus, il y a une tendance à la partialité envers les documents plus longs, et la nature statique du TF-IDF signifie qu'il ne se met pas à jour dynamiquement avec de nouveaux documents ou des changements dans l'usage de la langue, réduisant potentiellement son efficacité avec le temps.

3.2.2 Sentence-BERT

3.2.2.1 BERT

BERT, ou *Bidirectional Encoder Representations from Transformers*, est une méthode révolutionnaire introduite par Devlin et al. [2019] pour la représentation des langues naturelles. Contrairement aux modèles de langue précédents qui traitaient les entrées textuelles de manière séquentielle, soit de gauche à droite, soit de droite à gauche, BERT utilise le mécanisme d'attention de l'architecture des transformeurs pour prendre en compte simultanément le contexte des deux côtés d'un token dans une phrase. BERT est conçu pour pré-entraîner des représentations bidirectionnelles profondes en tenant compte conjointement des contextes gauche et droit à travers toutes ses couches, ce qui en fait un modèle profondément bidirectionnel.

BERT surmonte les limitations des modèles pré-entraînés antérieurs, tels que Word2Vec ou GloVe, qui généraient un seul vecteur de plongement pour chaque mot du vocabulaire, sans tenir compte du contexte spécifique dans lequel le mot apparaît. Ces modèles unidirectionnels ou bidirectionnels superficiels ne pouvaient pas capturer toute la complexité contextuelle de la langue naturelle. En revanche, la bidirectionnalité profonde de BERT lui permet de mieux comprendre le contexte global d'une phrase, améliorant ainsi, améliorant ainsi ses performances dans diverses tâches de TALN telles que la réponse aux questions, l'inférence linguistique et l'analyse des sentiments.

Au cœur de BERT se trouve un encodeur transformeur bidirectionnel multicouche, chaque couche comprenant deux composants principaux : un mécanisme d'auto-attention multi-têtes et un réseau entièrement connecté de type feed-forward. Le modèle est pré-entraîné en utilisant deux tâches non supervisées : la modélisation de langue masquée (*Masked Language Modeling*, MLM) et la prédiction de la prochaine phrase (*Next Sentence Prediction*NSP). Dans le MLM, BERT prédit des mots masqués aléatoirement en fonction de leur contexte, tandis que dans NSP, il détermine si une phrase suit logiquement une autre. Contrairement aux plongements

statiques, les plongements générés par BERT sont contextuels, permettant au modèle de gérer efficacement des phénomènes comme la polysémie.

3.2.2.2 Sentence-BERT

Traditionnellement, les plongements BERT pour les phrases et les documents étaient obtenus en prenant la sortie du modèle BERT pour chaque token et en appliquant une stratégie de pooling pour dériver un vecteur de longueur fixe pour l'ensemble de la phrase ou du document. Cependant, cette méthode échouait souvent à capturer efficacement les nuances sémantiques des textes plus longs [Reimers and Gurevych, 2019; Ferret, 2021]. Dans ce contexte, Reimers and Gurevych [2019] ont introduit Sentence-BERT (SBERT), une adaptation de l'architecture BERT qui utilise des structures de réseau siamoises et triplet pour générer des plongements de phrases sémantiquement significatifs plus efficacement.

En pratique, SBERT affine un modèle BERT pré-entraîné sur des tâches spécifiques impliquant des paires ou des triplets de phrases avec l'objectif de produire des plongements qui rapprochent sémantiquement des phrases similaires dans l'espace vectoriel tout en éloignant celles qui sont différentes. Cet ajustement est réalisée grâce à des objectifs d'entraînement tels que la similarité cosinus, la distance de Manhattan ou la distance euclidienne.

Plusieurs modèles SBERT pré-entraînés sont disponibles, adaptés à différents types de données, langues ou domaines. Ces modèles facilitent des applications telles que la similarité de phrases, le clustering et la recherche d'informations, offrant une réduction substantielle des ressources computationnelles et du temps nécessaires pour entraîner un modèle à partir de zéro.

3.3 Algorithmes de détection d'anomalies

Suite à notre discussion sur les techniques de représentation de texte dans la première section, nous nous tournons maintenant vers le deuxième composant : les algorithmes de détection d'anomalies.

La détection d'anomalies est un domaine de recherche très actif ces dernières années, avec une multitude d'algorithmes proposés, allant des approches d'apprentissage automatique traditionnelles aux méthodes d'apprentissage profond. Ces algorithmes utilisent divers mécanismes pour calculer les scores d'anomalie et sont appliqués sous différents paradigmes d'apprentissage. La tendance récente de la recherche est marquée par un changement de paradigme, passant de l'apprentissage non supervisé à l'apprentissage semi-supervisé ou faiblement supervisé, parallèlement au développement de nouveaux mécanismes de calcul des scores d'anomalie.

Compte tenu des défis spécifiques de la détection d'anomalies dans les textes, où des représentations vectorielles de haute dimension sont impliquées, notre stratégie de sélection et d'adaptation des algorithmes repose sur les principes suivants :

1. Utilisation d'algorithmes qui ont été testés sur des données tabulaires et d'image, ou qui sont spécifiquement conçus pour les données de haute dimension.
2. Sélection d'algorithmes qui n'ont pas été systématiquement évalués sur des données textuelles.

3. Couverture d'un large éventail de mécanismes traditionnels de calcul des scores d'anomalie, y compris les méthodes basées sur la proximité, les statistiques-probabilités, la reconstruction et l'apprentissage ensembliste. Nous incluons également des mécanismes récemment proposés, tels que l'apprentissage de représentation d'anomalies et l'apprentissage de scores d'anomalie, afin d'évaluer leurs avantages relatifs.
4. Pour les mécanismes traditionnels, intégration de méthodes avec des approches innovantes, telles que les algorithmes basés sur les angles ou les ordres pour les méthodes à base de proximité, et les ensembles hétérogènes pour l'apprentissage ensembliste.
5. Pour chaque type de mécanisme, évaluation à la fois des méthodes d'apprentissage profond et des méthodes d'apprentissage automatique traditionnelles afin d'examiner la nécessité et l'efficacité de l'inclusion des réseaux de neurones.

En appliquant ces critères, nous avons examiné sept algorithmes dans cette section.

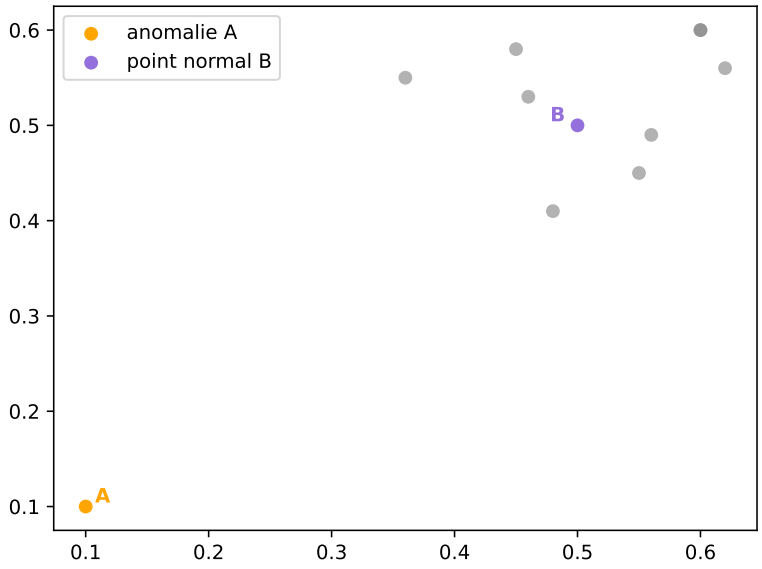
3.3.1 ABOD

ABOD (*Angle-Based Outlier Detection*) [Kriegel et al., 2008] est un modèle de détection d'anomalies basé sur la proximité, particulièrement conçu pour les données de haute dimension. La majorité des méthodes basées sur la proximité reposent sur la mesure des distances dans un espace euclidien à toutes les dimensions. Toutefois, dans les espaces à haute dimension, les distances peuvent perdre de leur pertinence à mesure que la dimensionnalité augmente, ce qui est connu sous le nom de « malédiction de la dimensionnalité ». ABOD est proposé pour résoudre ce problème en se concentrant sur les relations angulaires plutôt que sur les distances.

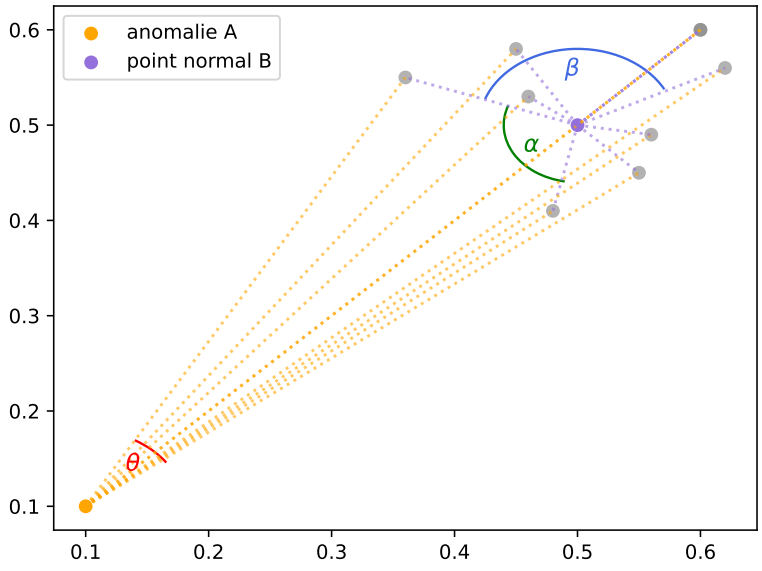
L'idée centrale des méthodes basées sur les angles est que les points anormaux, généralement situés en dehors du cluster principal de données, ont tendance à former des angles plus petits avec d'autres points de données. En revanche, les points bien intégrés dans le cluster de données sont entourés d'autres points couvrant diverses directions, ce qui se traduit par une plus grande variété d'angles.

Considérons un jeu de données comme illustré dans la Figure 3.1, où le point B est situé au centre du cluster et le point A est une anomalie située loin du cluster. Tous les autres points de données forment un angle (θ) limité autour du point A, indiquant son statut d'anomalie. D'autre part, pour les points à l'intérieur du cluster comme B, les angles entre les paires de points de données (par example, α et β) varient considérablement. En effet, plus un point de données est isolé des autres points, plus l'angle sous-jacent est vraisemblablement petit. Les points de données avec un spectre des angles plus petit sont des anomalies, tandis que ceux avec un spectre des angles plus large sont normaux. Ainsi, la variance dans le spectre des angles peut servir d'indicateur fiable d'anomalie au sein d'un ensemble de données.

Formellement, soit un jeu de données $\mathcal{D}$ avec un point de requête $\vec{A} \in \mathcal{D}$ et toute paire de points $\vec{B}, \vec{C} \in \mathcal{D}$, le vecteur $\overline{AB}$ représente la différence $\vec{B} - \vec{A}$. Pour un point anormal $\vec{A}$, les angles entre les vecteurs $\overline{AB}$ et $\overline{AC}$ ont tendance à montrer peu de variation entre les différentes paires $\vec{B}$ et $\vec{C}$. Ces angles sont ensuite inversement pondérés par la distance entre les points pour atténuer l'influence des paires distantes, réduisant ainsi efficacement les angles pondérés pour les points aberrants. Cet ajustement impacte ultérieurement le spectre global des angles associés au point de requête.



(a)



(b)

FIGURE 3.1 – ABOD (*Angle-Based Outlier Detection*)

Le facteur d'anomalie basé sur l'angle (*Angle-Based Outlier Factor*, ABOF) pour un point $\vec{A}$ est ainsi défini comme la variance de ces angles pondérés sur toutes les paires de points dans $\mathcal{D}$:

$$\textit{ABOF}(\vec{A}) = \textit{VAR}_{\vec{B},\vec{C}\in\mathcal{D}} \left(\frac{\langle \overline{AB}, \overline{AC} \rangle}{\|\overline{AB}\|_2^2 \cdot \|\overline{AC}\|_2^2} \right)$$

Ici, $\|\cdot\|^2$ représente la norme euclidienne au carré (L_2-norme), et $\langle\cdot\rangle$ désigne le produit scalaire. L'angle entre deux vecteurs de différence quelconques est normalisé par le produit de leurs longueurs au carré, ajustant ainsi l'effet de la distance dans le calcul de l'angle.

L'approche ABOD de base prend en compte tous les points de données, ce qui entraîne une grande complexité de calcul. Sa complexité temporelle de $O(n^3)$ la rend moins intéressante par rapport à des méthodes plus efficaces comme LOF, qui fonctionne en $O(n^2 \cdot k)$. Pour accélérer l'approche, FastABOD a été développé pour améliorer l'efficacité par approximation.

Contrairement à l'ABOD de base, FastABOD applique une technique d'élagage à base de bornes pour réduire la complexité de calcul en établissant des limites. Cette approche se concentre sur le calcul des angles uniquement entre un point et ses k plus proches voisins, qui sont les plus influents pour déterminer la variance des angles. Le principe sous-jacent est que les points éloignés les uns des autres ont moins d'impact sur le ABOF, ce qui rend leur prise en compte souvent inutile pour une détection précise des anomalies.

L'ABOF approximatif est défini comme suit :

$$\textit{approxABOF}_k(\vec{A}) = \textit{VAR}_{\vec{B},\vec{C}\in\mathcal{N}_k(\vec{A})} \left(\frac{\langle \overline{AB}, \overline{AC} \rangle}{\|\overline{AB}\|_2^2 \cdot \|\overline{AC}\|_2^2} \right)$$

où $\mathcal{N}_k(\vec{A}) \subseteq D$ représente les k plus proches voisins du point $\vec{A}$.

En se concentrant sur les points les plus pertinents, FastABOD échange un certain degré de précision contre des gains significatifs en vitesse et en efficacité computationnelle. Cette méthode optimisée aboutit à une complexité temporelle de $O(n^2+n\cdot k^2)$, rendant FastABOD particulièrement adapté aux grands jeux de données ou aux applications nécessitant un traitement en temps réel.

3.3.2 COPOD

COPOD (*Copula-Based Outlier Detection*) [Li et al., 2020b] est un modèle de détection d'anomalies basé sur la probabilité, conçu pour remédier au problème de la complexité de calcul élevée et la faible interprétabilité des méthodes traditionnelles, telles que l'OCSVM et les GMMs, lors du traitement de données à haute dimension.

COPOD construit une copule empirique et prédit ensuite les probabilités de queue pour chaque point de données. Ces probabilités de queue déterminent le niveau d'« extrémalité » du point de données. Une copule est une fonction de distribution cumulative (FDC) multivariée qui décrit la dépendance (inter-corrélation) entre des variables aléatoires. Elle permet de séparer les distributions marginales (comportements des variables individuelles) de la structure de dépendance (comment les variables sont liées les unes aux autres).

Étant donné un vecteur aléatoire d-dimensionnel $(X_1, X_2, \ldots, X_d)$, où les FDC marginales $F_j(x) = \mathbb{P}(X_j \leq x)$ sont des fonctions continues. En appliquant la transformation intégrale de probabilité à chaque composante, nous rendons les marginales uniformément distribuées sur l'intervalle $[0, 1]$:

$$(U_1, U_2, \ldots, U_d) = (F_1(X_1), F_2(X_2), \ldots, F_d(X_d))$$

La copule C de $\boldsymbol{X}$ est définie comme la FDC jointe de $\boldsymbol{U}$:

$$C(u_1, u_2, \ldots, u_d) = \mathbb{P}(U_1 \leq u_1, U_2 \leq u_2, \ldots, U_d \leq u_d)$$

où u_i représente la valeur marginale uniforme pour la i-ème composante.

Le théorème de Sklar [Sklar, 1959] indique que toute distribution jointe multivariée $F(x_1, \ldots, x_d)$ peut être exprimée en termes de fonctions de distribution marginales univariées $F_1, \ldots, F_d$ et d'une copule :

$$F(\boldsymbol{x}) = C(F_1(x_1), \ldots, F_d(x_d))$$

Ainsi, une copule nous permet de décrire la distribution jointe de $(X_1, \ldots, X_d)$ en utilisant uniquement leurs marginales. Cette méthode est très pratique pour traiter les données de haute dimension, car elle permet de modéliser chaque dimension de manière indépendante tout en fournissant un mécanisme fiable pour combiner ces distributions marginales en une distribution jointe complète.

Pour un jeu de données d-dimensionnel avec n échantillons $\boldsymbol{X} = (X_{1,i}, \ldots, X_{d,i}), i = 1, \ldots, n$, le processus de détection d'anomalies avec COPOD se déroule en trois étapes :

Étape 1 : Calcul des FDC empiriques. Calculer les FDC empiriques pour chaque variable (caractéristique) dans le jeu de données. La FDC empirique $\hat{F}(x)$ est définie comme :

$$\hat{F}(x) = \mathbb{P}((-\infty, x]) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_i \leq x)$$

où $\mathbb{I}(\cdot)$ est la fonction indicatrice qui vaut 1 lorsque son argument est vrai et 0 sinon.

Étape 2 : Construction de la copule empirique. Construire une copule empirique basée sur les FDC empiriques. D'abord, les observations de la copule empirique sont obtenues par transformation uniforme :

$$(\hat{U}_{1,i}, \ldots, \hat{U}_{d,i}) = (\hat{F}(X_{1,i}), \ldots, \hat{F}(X_{d,i}))$$

Ensuite, en incorporant les observations de la copule empirique dans la fonction copule, COPOD construit la copule empirique :

$$\hat{C}(u_1, \ldots, u_d) = \frac{1}{n} \sum_{i=1}^{n} \prod \mathbb{I}(\tilde{U}_{1,i} \leq u_1, \ldots, \tilde{U}_{d,i} \leq u_d)$$

où $\tilde{U}_{j,i}$ représente les probabilités de queue droite pour chaque dimension j et chaque observation i, autrement dit $\tilde{U}_{j,i} = 1 - \hat{F}_j(X_{j,i})$.

Étape 3 : Prédiction de la probabilité de queue. Estimer les probabilités de queue pour chaque observation x_i en utilisant la copule empirique. COPOD considère les anomalies comme des événements de queue. Cette étape a donc pour objectif de calculer les probabilités de queue pour chaque observation x_i :

— Probabilité de queue gauche : $F_X(x_i) = \mathbb{P}(X_1 \leq x_{1,i}, \ldots, X_d \leq x_{d,i})$

— Probabilité de queue droite : $1 - F_X(x_i) = \mathbb{P}(X_1 > x_{1,i}, \ldots, X_d \geq x_{d,i})$

Exprimées en termes de copule empirique :

$$\hat{F}_X(x_i) = \hat{C}(\hat{U}_{1,i}, \ldots, \hat{U}_{d,i}) = \frac{1}{n} \sum_{i=1}^{n} \prod_{j=1}^{d} \mathbb{I}(\tilde{U}_{j,i} \leq u_j)$$

$$1 - \hat{F}_X(x_i) = 1 - \hat{C}(\hat{U}_{1,i}, \ldots, \hat{U}_{d,i}) = \frac{1}{n} \sum_{i=1}^{n} \prod_{j=1}^{d} \mathbb{I}(\tilde{U}_{j,i} \geq u_j)$$

La probabilité de queue représente la chance de rencontrer un point aussi extrême que ou plus extrême que x_i. Une anomalie est identifiée lorsque la probabilité de queue gauche ou droite est faible, suggérant qu'il est peu probable d'observer un point de données avec une extrémalité plus élevée. Le score d'anomalie est calculé comme le logarithme négatif de la probabilité de queue. Une probabilité de queue plus faible se traduit par un score d'anomalie plus élevé, indiquant que l'observation est plus susceptible d'être une anomalie en fonction de son extrémité relatif à l'ensemble de données. Ce score représente une mesure relative de la vraisemblance que X_i soit une valeur aberrante, plutôt que sa probabilité absolue.

COPOD est une méthode non supervisée qui ne nécessite pas de données étiquetées lors de l'apprentissage, ce qui la rend extrêmement flexible et accessible. Sa nature déterministe signifie qu'elle fonctionne tout simplement sans hyperparamètres, ce qui élimine la complexité du tuning tout en garantissant la reproductibilité et la cohérence de ses résultats. L'efficacité de COPOD en fait un choix idéal pour les jeux de données de haute dimension, offrant des performances robustes sans sacrifier la vitesse, répondant ainsi aux exigences des scénarios de données à grande échelle. En outre, COPOD se distingue par son interprétabilité ; il quantifie la contribution d'anormalité de chaque dimension grâce au graphe dimensionnel des valeurs aberrantes, offrant ainsi un aperçu clair de la structure des données.

3.3.3 ECOD

ECOD (Empirical Cumulative Distribution-based Outlier Detection) [Li et al., 2022c] est un autre modèle de détection d'anomalies peu profond basé sur la probabilité, spécialement conçu pour les données de grande taille et à haute dimension. De même que le modèle COPOD, ECOD considère les anomalies comme des événements de queue, qui sont des occurrences rares dans les régions de faible densité de la distribution de probabilité. L'idée fondamentale derrière ECOD est d'estimer la FDC empirique des données, puis de dériver les scores d'anomalie à partir des probabilités de queue.

Les données à haute dimension posent un problème important lors de l'application des FDC empiriques traditionnelles. À mesure que le nombre de dimensions augmente, la FDC empirique jointe sur toutes les variables converge plus difficilement vers la véritable FDC jointe. Tout comme COPOD, ECOD est spécialement conçu pour résoudre cette malédiction de la dimensionnalité.

COPOD utilise le théorème de Sklar pour calculer séparément les distributions marginales de chaque dimension et utilise des copules pour représenter la FDC jointe. Cette approche permet de modéliser les dépendances entre les dimensions en utilisant des copules, qui séparent les distributions marginales de la structure de dépendance d'une distribution multivariée.

ECOD, en revanche, fonctionne sous l'hypothèse simplificatrice d'indépendance entre les dimensions. Soit un jeu de données $\mathbf{X}$ à d dimensions avec n observations. Nous utilisons $X_{i,j}$ pour désigner la i-ème observation de la j-ème dimension. Cette hypothèse peut être représentée mathématiquement comme suit :

$$F(X_1, X_2, \ldots, X_d) = \prod_{j=1}^{d} F_j(X_j)$$

où $F(X_1, X_2, \ldots, X_d)$ est la FDC jointe pour le vecteur $\mathbf{X}_i \in \mathbb{R}^d$ et $F_j(x) = \mathbb{P}(X_j \leq x), x \in \mathbb{R}$ représente la FDC univariée pour la dimension j. Bien que cette hypothèse ne soit pas forcément toujours vraie, ECOD a démontré son efficacité dans de nombreux scénarios pratiques, y compris les cas où les caractéristiques sont interdépendantes.

La méthode ECOD fonctionne en deux étapes principales :

Étape 1 : Calcul des FDC univariées et des probabilités de queue. La FDC univariée est estimée en utilisant la FDC empirique :

$$\hat{F}_j^{gauche}(z) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_{i,j} \leq z), \quad z \in \mathbb{R}$$

La FDC de la queue droite est :

$$\hat{F}_j^{droite}(z) = 1 - \hat{F}_j^{gauche}(z) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_{i,j} \geq z), \quad z \in \mathbb{R}$$

Ainsi, nous pouvons estimer les FDC empiriques jointes des queues gauche et droite sur toutes les d dimensions sous l'hypothèse d'indépendance :

$$\hat{F}^{gauche}(x) = \prod_{j=1}^{d} \hat{F}_j^{gauche}(x_j) \quad \text{et} \quad \hat{F}^{droite}(x) = \prod_{j=1}^{d} \hat{F}_j^{droite}(x_j)$$

Étape 2 : Agrégation des scores d'anomalie. Pour chaque point de données $\mathbf{X}_i$, ses probabilités de queue sont agrégées pour former un score d'anomalie final $S_i \in (0, \infty]$. Pour mesurer le degré d'anomalie d'un point de données, ECOD calcule ses probabilités de queue dans toutes les dimensions, en supposant qu'elles sont indépendantes. Cette étape demande un choix de la probabilité de queue à utiliser, soit à droite, soit à gauche. La décision est prise en fonction de l'asymétrie du jeu des données. Le coefficient d'asymétrie pour la dimension j est défini comme suit :

$$\gamma_j = \frac{\frac{1}{n} \sum_{i=1}^{n} (X_{i,j} - \overline{X}_j)^3}{\left[\frac{1}{n-1} \sum_{i=1}^{n} (X_{i,j} - \overline{X}_j)^2\right]^{3/2}}$$

où $\overline{X}_j$ est la moyenne de la dimension j.

En ce qui concerne le score d'anomalie final, ECOD fonctionne dans l'espace des probabilités logarithmiques négatives comme COPOD. La méthode implique de calculer trois scores différents et de sélectionner la valeur maximale parmi eux :

1. **Score basé sur la queue gauche (S_i^{gauche}) :**

$$S_i^{gauche}(X_i) = -\sum_{j=1}^{d} \log\Big(\hat{F}_j^{gauche}(X_{i,j})\Big)$$

2. **Score basé sur la queue droite (S_i^{droite}) :**

$$S_i^{droite}(X_i) = -\sum_{j=1}^{d} \log\Big(\hat{F}_j^{droite}(X_{i,j})\Big)$$

3. **Score automatiquement choisi (S_i^{auto}) :**

$$S_i^{auto}(X_i) = -\sum_{j=1}^{d} \Big[\mathbb{I}(\gamma_j < 0) \log\Big(\hat{F}_j^{gauche}(X_{i,j})\Big) + \mathbb{I}(\gamma_j \geq 0) \log\Big(\hat{F}_j^{droite}(X_{i,j})\Big)\Big]$$

Le score d'anomalie final S_i pour chaque point de données $\mathbf{X}_i$ est :

$$S_i = \max\{S_i^{gauche}(X_i), S_i^{droite}(X_i), S_i^{auto}(X_i)\}$$

ECOD offre un cadre efficace pour détecter les anomalies dans les données à haute dimension en utilisant des distributions cumulatives empiriques et des probabilités de queue. Sa gestion de la malédiction de la dimensionnalité et son calcul efficace en font un outil précieux en détection d'anomalies. Cependant, l'hypothèse d'indépendance sur laquelle repose ECOD n'est pas toujours vérifiée dans les données textuelles transformées. Malgré cela, ses performances prouvées dans d'autres contextes avec des caractéristiques interdépendantes suggèrent qu'il serait intéressant de tester ECOD pour la détection d'anomalies textuelles. Cela pourrait valider son efficacité et offrir de nouvelles perspectives d'application.

3.3.4 ALAD

ALAD (*Adversarially Learned Anomaly Detection*) [Zenati et al., 2018] est une méthode semi-supervisée utilisant les réseaux antagonistes génératifs (*Generative Adversarial Networks*, GANs), qui sont reconnus pour leur capacité à modéliser des distributions de données complexes et de haute dimension. Un GAN typique comprend deux composants principaux : un générateur (G) et un discriminateur (D). Le générateur est entraîné pour produire des données qui ressemblent aux données réelles, tandis que le discriminateur est entraîné pour différencier les données générées des données réelles. Dans le contexte de la détection d'anomalies, un GAN est généralement entraîné sur des échantillons de données normaux pour apprendre un mappage de la représentation de l'espace latent aux échantillons réalistes. Ce mappage est ensuite utilisé pour projeter de nouveaux échantillons non vus dans l'espace latent. Le générateur, ayant été entraîné exclusivement sur des données normales, tend à reconstruire toute entrée anormale comme normale. Par conséquent, l'écart entre l'entrée et sa reconstruction peut être utilisé pour identifier les anomalies.

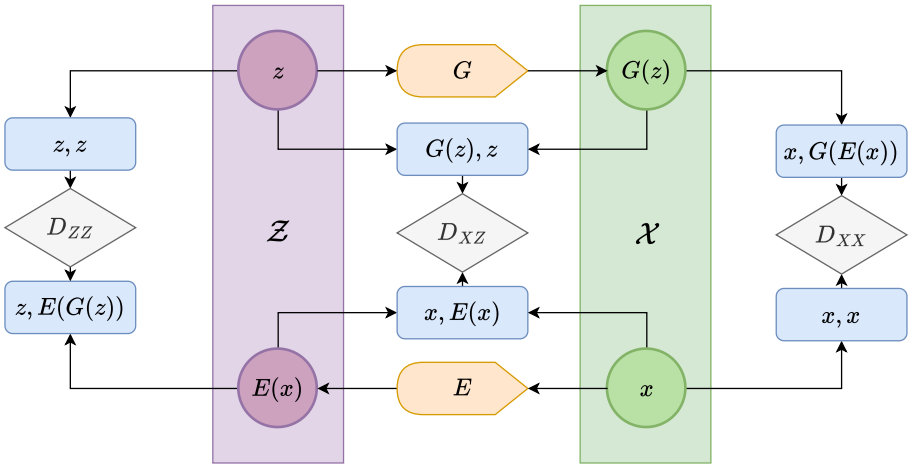


FIGURE 3.2 – ALAD (*Adversarially Learned Anomaly Detection*)

Contrairement aux méthodes basées sur les GANs standard comme AnoGAN, ALAD intègre une architecture GAN bidirectionnelle en ajoutant un encodeur supplémentaire (E). Grâce à cet encodeur, l'ALAD établit également une correspondance entre les échantillons de données et les variables latentes, ce qui lui permet d'éviter la procédure d'inférence coûteuse en calcul requise par d'autres méthodes (par exemple, l'AnoGAN) lors de la récupération des variables latentes pour la détection d'anomalies.

L'architecture d'ALAD comprend plusieurs composants (voir Figure 3.2) :

— Le générateur (G) crée des données synthétiques $G(z)$ à partir de variables latentes aléatoires z dans l'espace latent $\mathcal{Z}$.

— Le codeur (E) met en correspondance les données originales $\mathcal{X}$ avec l'espace latent $\mathcal{Z}$, dans le but de créer une représentation latente (approximative) $E(x)$ pour un point de données x. Un discriminant supplémentaire D_{XX} est ajouté pour garantir la cohérence du cycle, c'est-à-dire que $G(E(x)) \approx x$.

— Le discriminateur D_{ZZ} compare les variables latentes réelles z aux variables latentes encodées $E(G(z))$ pour s'assurer que les échantillons générés ont des représentations latentes similaires à celles des données réelles.

— Le discriminateur D_{XZ} compare les paires de données réelles x et leurs versions encodées $E(x)$ avec les données générées $G(z)$ et leurs encodages z pour s'assurer de la cohérence des processus d'encodage et de génération.

— Le discriminateur D_{XX} compare les paires de données réelles (x, x) avec les paires générées et recodées $(x, G(E(x))$ pour assurer l'exactitude du processus de reconstruction. L'erreur de reconstruction est utilisée comme score d'anomalie. Le générateur, le discrminateur et l'encodeur sont formés conjointement sur des données normales.

Le modèle ALAD, initialement conçu pour des données complexes et de haute dimension comme les images, a été efficacement appliqué à des ensembles de données d'images et tabulaires. Pour adapter l'ALAD aux données textuelles, il est crucial de représenter et de prétraiter correctement le texte.

3.3.5 XGBOD

XGBOD (*Extreme Boosting Based Outlier Detection*) [Zhao and Hryniewicki, 2018] est un modèle ensembliste qui intègre un processus d'apprentissage de caractéristiques pour la détection des anomalies. Il s'agit d'une approche hybride combinant des techniques supervisées et non supervisées. XGBOD extrait des représentations utiles des données en utilisant plusieurs algorithmes de détection d'anomalies non supervisés, augmentant ainsi les capacités prédictives d'un classificateur supervisé intégré.

Ce modèle est applicable aussi bien dans un scénario faiblement supervisé que dans un scénario entièrement supervisé. Il est particulièrement adapté aux situations où les anomalies sont connues mais rares, car il peut utiliser efficacement un nombre limité de données étiquetées. Même une petite quantité de données étiquetées peut considérablement améliorer les performances de XGBOD dans des contextes faiblement supervisés.

XGBOD opère en trois phases distinctes :

Phase 1 : Apprentissage de représentation non supervisé (génération de TOS). Dans cette phase initiale, diverses méthodes de détection d'anomalies non supervisées telles que KNN, AvgKNN, LOF, IForest, HBOS et OCSVM sont appliquées aux données originales. Ces méthodes génèrent des scores d'anomalie transformés (*Transformed Outlier Scores*, TOS), qui servent de représentations de données supplémentaires capturant des informations liées aux anomalies.

Soit $\mathbf{X} \in \mathbb{R}^{n \times d}$ le jeu de données originales avec n points de données et d caractéristiques. Les fonctions de score d'anomalie, $\phi_i(\cdot)$, génèrent une matrice de scores d'anomalie $\phi(\mathbf{X}) \in \mathbb{R}^{n \times k}$, où k représente le nombre de fonctions de score d'anomalie utilisées.

$$\phi(\mathbf{X}) = [\phi_1(\mathbf{X}), \phi_2(\mathbf{X}), \ldots, \phi_k(\mathbf{X})]^T$$

Phase 2 : Augmentation des caractéristiques. La deuxième phase implique la sélection d'un sous-ensemble de TOS en utilisant plusieurs méthodes de séléction. Les TOS choisis sont ensuite concaténés avec les caractéristiques originales pour créer un espace de caractéristiques augmenté. Cet espace de caractéristiques enrichi intègre les caractéristiques des données originales avec les indicateurs d'anomalies nouvellement dérivés, fournissant ainsi une représentation des données plus complète pour les analyses subséquentes. Les TOS sont sélectionnées selon 3 stratégies (voir Figure 3.3) :

- **Sélection aléatoire** : Choisir p TOS de manière aléatoire parmi les k TOS disponibles.
- **Sélection précise** : Sélectionner les p TOS les plus précis en utilisant une métrique d'évaluation telle que le score AUCROC.
- **Sélection équilibrée** : Maintenir un équilibre entre exactitude et diversité en choisissant les TOS avec la meilleure exactitude pondérée par leur diversité. La fonction de exactitude pondérée est définie par :

$$\psi(\phi_i) = \frac{ACC(\phi_i)}{1 + \sum_{j \in S} \rho(\phi_i, \phi_j)}$$

où $ACC(\phi_i)$ est l'exactitude de ϕ_i et $\rho(\phi_i, \phi_j)$ est la corrélation de Pearson entre ϕ_i et ϕ_j.

Phase 3 : Classificateur XGBoost. La phase finale utilise un classificateur *Extreme Gradient Boosting* (XGBoost), entraîné sur l'espace de caractéristiques augmenté. Le gradient boosting est une technique d'ensemble qui emploie de multiples arbres de décision comme apprenants faibles pour former un modèle prédictif robuste. Chaque arbre est construit séquentiellement pour corriger les erreurs des arbres précédents, et XGBoost améliore ce processus en intégrant des termes de régularisation pour prévenir le surapprentissage et optimiser à la fois la vitesse de calcul et la performance du modèle.

XGBOD a montré une performance supérieure à travers divers benchmarks de détection d'anomalies, surpassant les détecteurs individuels, les ensembles complets et d'autres algorithmes basés sur l'apprentissage de représentation sur plusieurs jeux de données d'anomalies. Initialement conçu pour et largement testé sur des données tabulaires, XGBOD fait face à un principal défi lorsqu'il est adapté aux données textuelles : la représentation efficace du texte pour la détection d'anomalies. Cela nécessite de transformer le texte non structuré en un format structuré qui s'aligne sur les exigences de la méthodologie XGBOD, garantissant ainsi que les données textuelles peuvent être efficacement incorporées dans le processus de détection d'anomalies.

3.3.6 DevNet

DevNet (Deviation Networks) [Pang et al., 2019] est un modèle de détection d'anomalies profond, conçu pour répondre aux défis de la détection d'anomalies dans des données complexes et à haute dimension.

Les méthodes traditionnelles de détection d'anomalies profondes reposent généralement sur un processus en deux étapes : elles reconstruisent les données d'entrée (comme avec AnoGan et VAE) ou apprennent de nouvelles caractéristiques (comme avec REPEN et DeepSVDD), puis elles calculent les scores d'anomalie à partir de ces représentations reconstruites ou nouvellement générées.

Ces méthodes souffrent souvent d'inefficacité en raison de la nature indirecte de leurs mécanismes de calcul de score, ce qui peut nuire à l'exactitude de la détection. La dépendance à la reconstruction ou à la transformation des caractéristiques peut introduire du bruit et de la complexité, diluant potentiellement l'exactitude et l'efficacité de la détection des anomalies, surtout dans des ensembles de données à haute dimension où la distinction entre les instances normales et anormales peut être subtile et complexe. En revanche, DevNet adopte une approche unifiée qui apprend et optimise directement les scores d'anomalie à partir des données d'entrée grâce à un processus d'apprentissage de la déviation de bout en bout.

DevNet fonctionne sous un paradigme d'apprentissage semi-supervisé PU (*Positive-Unlabeled*) en utilisant un petit ensemble d'anomalies étiquetées pour guider efficacement son processus d'apprentissage. Contrairement aux méthodes non supervisées traditionnelles, qui souffrent souvent d'une inefficacité dans l'utilisation des données et d'un score de détection sous-optimal, DevNet tire parti de cette information partielle pour mieux distinguer les anomalies des données normales.

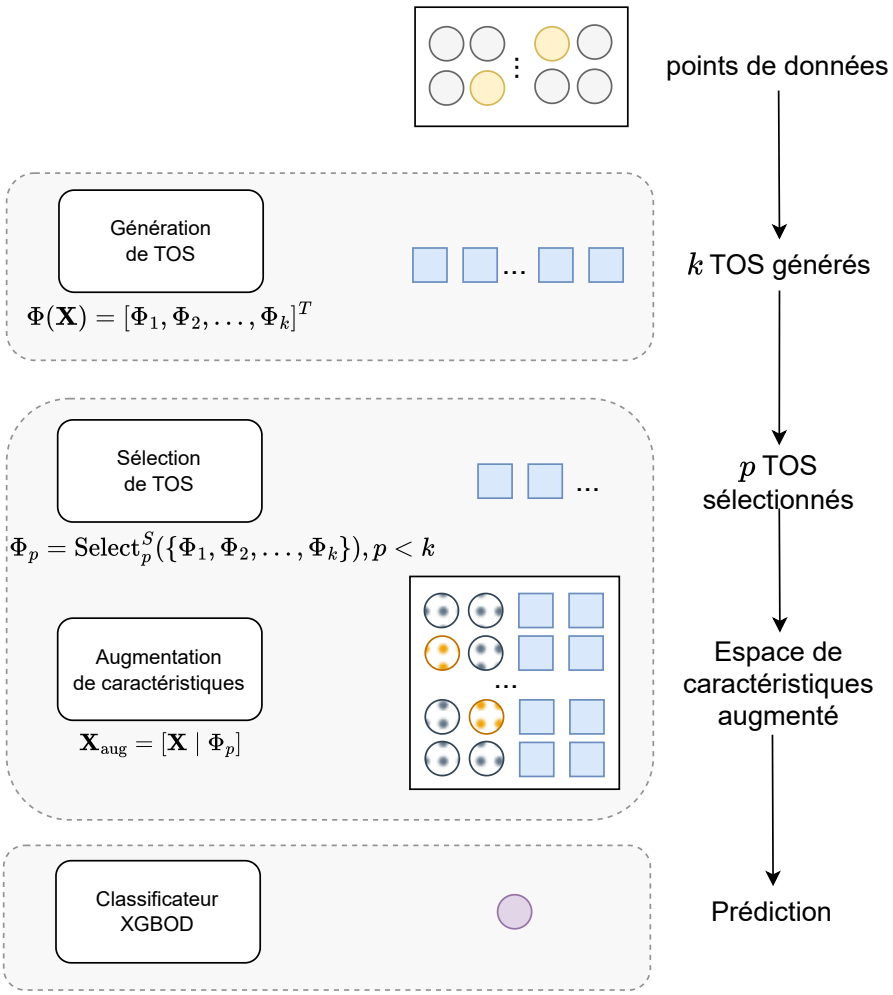


FIGURE 3.3 – L'algorithme XGBOD (*Extreme Boosting Based Outlier Detection*) en trois étapes principales : (1) **Génération de TOS** : Plusieurs méthodes de détection d'anomalies non supervisées génèrent des scores d'anomalies transformés (TOS) à partir du jeu de données original $\mathbf{X}$. (2) **Sélection de TOS et augmentation de caractéristiques** : Les TOS sélectionnés (ϕ_p) sont concaténés avec les caractéristiques originales pour créer un espace de caractéristiques augmenté ($\mathbf{X}_{\text{aug}} = [\mathbf{X} \mid \phi_p]$). Les stratégies de sélection incluent des approches aléatoires, précises et équilibrées. (3) **Classificateur XGBoost** : L'espace de caractéristiques augmenté est utilisé pour entraîner un classificateur XGBoost, qui prédit les anomalies en exploitant la représentation enrichie.

Aperçu du modèle L'objectif du modèle DevNet est d'apprendre une fonction de scoring d'anomalie $\phi : \mathcal{X} \to \mathbb{R}$, qui attribue des scores plus élevés aux anomalies qu'aux points de données normales au sein d'un ensemble de données $\mathcal{X} = \mathcal{A} \cup \mathcal{U}$, où $\mathcal{U}$ représente les données non étiquetées et $\mathcal{A}$ est un petit ensemble d'anomalies étiquetées avec $|\mathcal{A}| \ll |\mathcal{U}|$. Pour ce faire, DevNet introduit une distribution de probabilité a priori des scores d'anomalie et une nouvelle fonction de perte pour entraîner un détecteur d'anomalies profondes de bout en bout, avec l'objectif d'attribuer des scores d'anomalie statistiquement significativement plus élevés aux anomalies qu'aux objets normaux. Ainsi, le modèle se compose de trois éléments principaux (Figure 3.4) :

1. **Réseau de scoring d'anomalie (ϕ)** : Produit un score d'anomalie scalaire pour chaque entrée $\mathbf{x}$.
2. **Générateur de scores de référence** : Génère un score de référence $\mu_{\mathcal{R}}$ qui est la moyenne des scores d'anomalie pour un ensemble d'échantillons normaux sélectionnés aléatoirement. Ce score de référence est déterminé par une distribution de probabilité a priori, ce qui facilite la génération de scores d'anomalie efficaces et interprétables.
3. **Fonction de perte de déviation** : Utilise $\phi(x)$, $\mu_{\mathcal{R}}$ et l'écart-type $\sigma_{\mathcal{R}}$ pour guider l'optimisation des scores d'anomalie. Elle assure que les scores d'anomalie des anomalies s'écartent significativement de $\mu_{\mathcal{R}}$ dans la queue supérieure tandis que les scores des données normales sont proches de $\mu_{\mathcal{R}}$.

Réseau de scoring d'anomalie de bout en bout Soit $\mathcal{Q} \in \mathbb{R}^M$ un espace de représentation intermédiaire à M dimension. Un réseau de scoring d'anomalie ϕ peut être défini comme une combinaison de deux composants :

1. **Apprenant de représentation des caractéristiques (ψ)** : L'apprenant de caractéristiques $\psi(\cdot;\Theta_r) : \mathcal{X} \mapsto \mathcal{Q}$ est un réseau de neurones avec H couches cachées. Il mappe l'entrée $\mathbf{x}$ en une représentation intermédiaire $\mathbf{q}$:

$$\mathbf{q} = \psi(\mathbf{x};\Theta_r)$$

   où $\mathbf{x} \in \mathcal{X}$, $\mathbf{q} \in \mathcal{Q}$ et Θ_r représente les paramètres du réseau. La structure du réseau varie selon le type de données d'entrée, comme des réseaux convolutifs pour les données d'image et des réseaux récurrents pour les données séquentielles.
2. **Apprenant de score d'anomalie (η)** : L'apprenant de score d'anomalie $\eta(\cdot;\Theta_s) : \mathcal{Q} \mapsto \mathbb{R}$ est une seule couche linéaire qui mappe la représentation intermédiaire $\mathbf{q}$ à un score d'anomalie scalaire :

$$\eta(q;\Theta_s) = \sum_{i=1}^{M} w_i^o q_i + w_{M+1}^o$$

   où Θ_s consiste en les poids w_i^o et le biais w_{M+1}^o.

Ainsi, la fonction de scoring d'anomalie complète ϕ est représentée comme suit :

$$\phi(\mathbf{x};\Theta) = \eta(\psi(\mathbf{x};\Theta_r);\Theta_s)$$

Cette configuration mappe directement les données d'entrée à des scores d'anomalie scalaires et peut être entraînée de bout en bout.

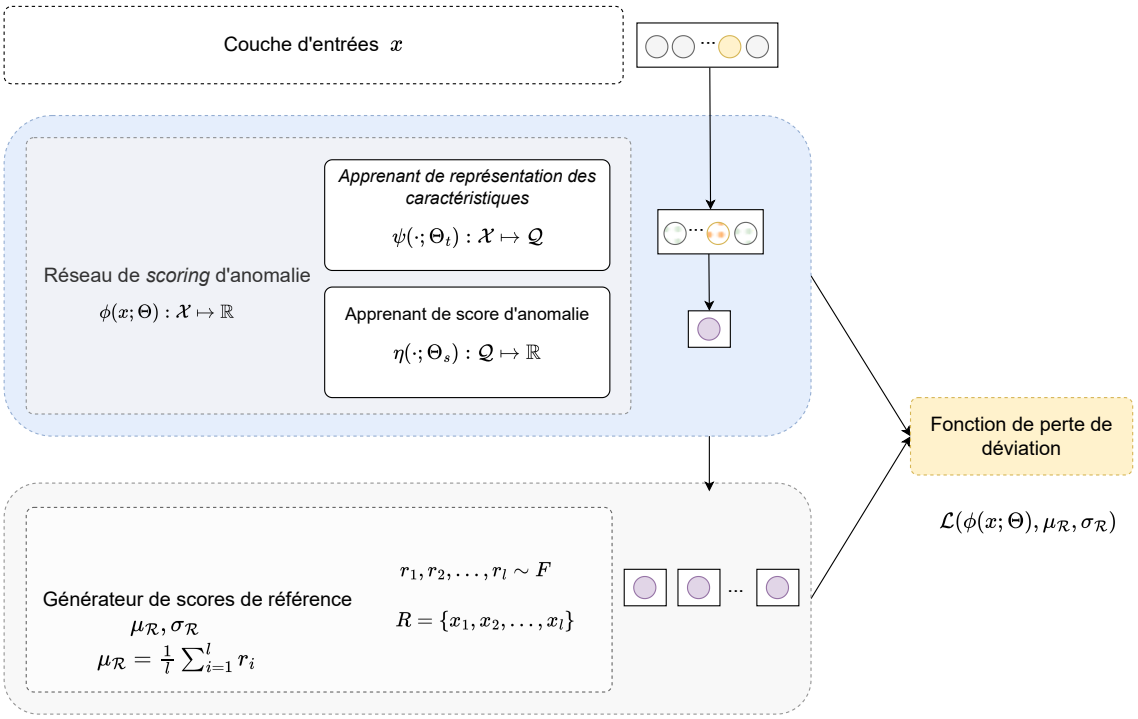


FIGURE 3.4 – Illustration du modèle DevNet. $\phi(x;\Theta)$ est un apprenant de score d'anomalie avec les paramètres Θ. $\mu_\mathcal{R}$ est la moyenne des scores d'anomalie de certains objets normaux, déterminée par une distribution de probabilité a priori F. $\sigma_\mathcal{R}$ est un écart-type associé à $\mu_\mathcal{R}$. La perte $\mathcal{L}(\phi(x;\Theta), \mu_\mathcal{R}, \sigma_\mathcal{R})$ est définie pour garantir que les scores d'anomalie des anomalies s'écartent de manière statistiquement significative de $\mu_\mathcal{R}$ dans la queue supérieure, tandis que les objets normaux ont des scores d'anomalie aussi proches que possible de $\mu_\mathcal{R}$.

Scores de référence basés sur une loi gaussienne Le score de référence μ_R est défini comme la moyenne des scores d'anomalie d'un ensemble de l échantillons normaux $\{r_1, r_2, \ldots, r_l\}$. En employant une loi gaussienne, les scores d'anomalie r_i sont tirés d'une distribution gaussienne :

$$r_1, r_2, \ldots, r_l \sim \mathcal{N}(\mu, \sigma^2)$$

$$\mu_\mathcal{R} = \frac{1}{l}\sum_{i=1}^{l} r_i$$

où μ et σ sont la moyenne et l'écart-type de la distribution gaussienne.

Perte de déviation basée sur le score Z La déviation d'un score d'anomalie par rapport au score de référence est calculée en utilisant le score Z :

$$dev(x) = \frac{\phi(x;\Theta) - \mu_\mathcal{R}}{\sigma_\mathcal{R}}$$

où $\sigma_\mathcal{R}$ est l'écart-type des scores de référence. La fonction de perte de déviation $\mathcal{L}$ est définie comme suit :

$$\mathcal{L}(\phi(x;\Theta), \mu_\mathcal{R}, \sigma_\mathcal{R}) = (1-y)|dev(\mathbf{x})| + y\max(0, a - dev(\mathbf{x}))$$

Ici, y est l'étiquette (1 pour les anomalies, 0 pour les données normales) et a est un paramètre d'intervalle de confiance. Cette perte assure que les scores d'anomalie pour les échantillons normaux restent aussi proches que possible de $\mu_{\mathcal{R}}$, tout en forçant une déviation minimale de a entre $\mu_{\mathcal{R}}$ et les scores d'anomalie des anomalies. Malgré l'absence de données normales étiquetées, DevNet a montré de bonnes performances en traitant les données non étiquetées dans $\mathcal{U}$ comme normales. Cette stratégie s'est avérée efficace même au sein de l'ensemble de données avec un ratio d'anomalie important [Pang et al., 2019].

DevNet est particulièrement efficace dans les scénarios avec un nombre limité d'anomalies étiquetées et des ensembles de données à haute dimension. Il surpasse de nombreuses méthodes existantes sur plusieurs benchmarks, excellant en termes de AUC-ROC et AUC-PR [Pang et al., 2019]. Cela en fait un candidat fort pour des applications telles que la détection d'anomalies dans les textes, où la complexité des données et la rareté des anomalies posent des défis significatifs.

3.3.7 PReNET

PReNET (*Pairwise Relation prediction-based ordinal regression NETwork*) [Pang et al., 2023] est un modèle d'apprentissage profond conçu pour la détection d'anomalies semi-supervisée [1]. Il introduit une nouvelle approche en convertissant le problème de la détection d'anomalies en une tâche de prédiction de la relation entre un couple d'observations.

Le modèle utilise un réseau neuronal de régression ordinale à deux flux (*two-stream ordinal regression network*) pour prédire les relations entre les paires d'observations. Ce réseau comporte deux branches parallèles qui apprennent les représentations des deux instances dans chaque paire. Ensuite, ces représentations sont combinées pour prédire une étiquette ordinale, reflétant la normalité relative des instances l'une par rapport à l'autre.

La régression ordinale permet de classer les paires d'instances selon un ordre spécifique, en attribuant des étiquettes ordinales qui indiquent la probabilité que les paires contiennent des anomalies. Cela signifie que PReNET identifie les anomalies en fonction de la position relative des instances anormales et normales dans l'espace de représentation appris, avec des paires contenant deux anomalies ayant des scores plus élevés que les paires avec une seule ou aucune anomalie.

Cette approche basée sur l'ordre permet à PReNET de généraliser plus efficacement, car elle exploite les différences relatives entre les instances normales et anormales plutôt que de se concentrer uniquement sur des anomalies spécifiques déjà observées.

Méthode semi-supervisée, PReNET fonctionne en deux phases (voir Figure 3.5) : l'entraînement et le test. Le processus d'apprentissage se compose de trois éléments consécutifs, alors que le processus de détection se concentre sur le calcul des scores d'anomalie.

1. L'article original considère PReNET comme une méthode faiblement supervisée en l'opposant aux méthodes semi-supervisées. Cette distinction est due au fait que l'article définit la semi-supervision dans un sens étroit et traditionnel, en se concentrant sur le paradigme d'apprentissage « *Normal-Only* (NO) ». Cependant, dans cette thèse, nous la considérons dans notre cadre théorique plus large de l'apprentissage semi-supervisé, qui comprend à la fois le paradigme d'apprentissage « Positive-Unlabeled (PU) » et celui d'apprentissage NO.

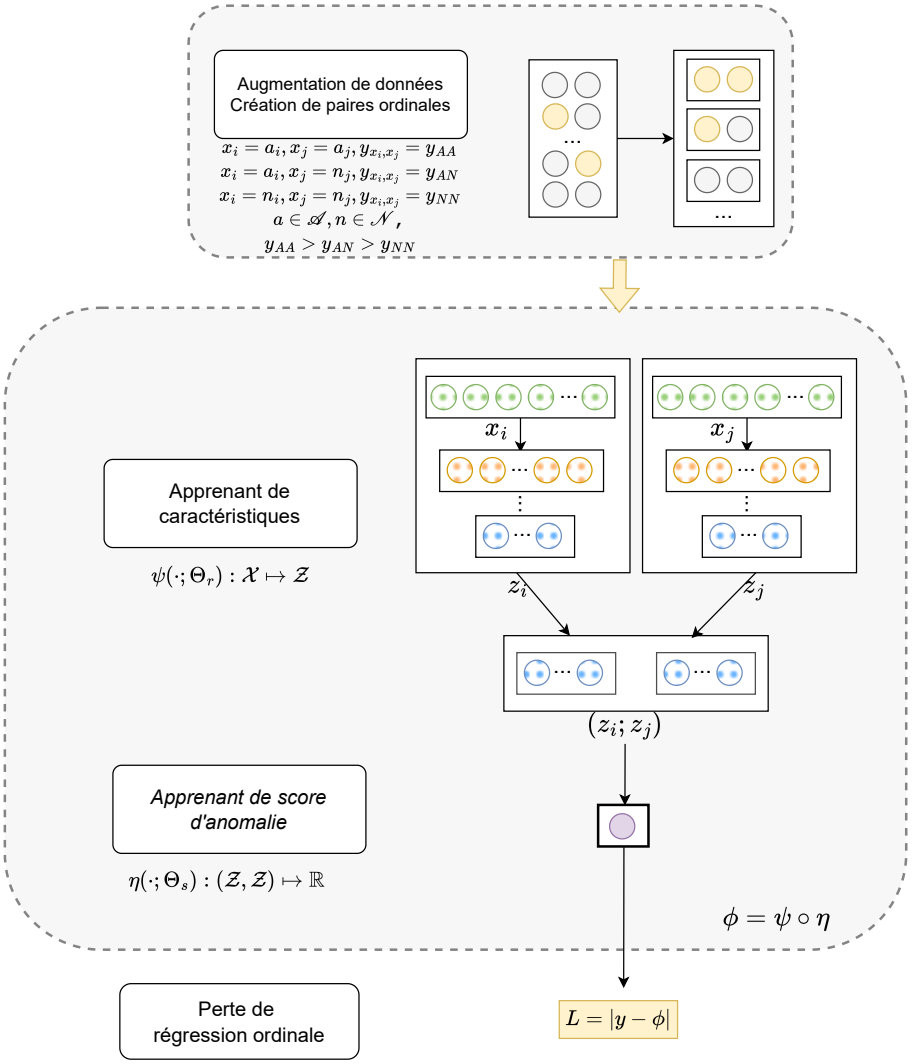


FIGURE 3.5 – L'algorithme PReNET (Pairwise Relation Prediction-based Ordinal Regression Network) en trois étapes principales : (1) **Apprentissage des relations de paire et augmentation des données** : Création de paires ordinales à partir de données anormales étiquetées et non étiquetées. (2) **Apprentissage de scores d'anomalies de bout en bout** : (i) **Apprenant de caractéristiques** : Transformation des instances de données en une représentation intermédiaire. (ii) **Apprenant de score d'anomalie** : Calcul des scores d'anomalies et (3)**Régression ordinale** : Minimisation de la perte de régression ordinale.

Apprentissage des relations de paire et augmentation des données PReNET utilise une méthode d'augmentation de données en deux étapes, basée sur l'appariement, afin d'élargir considérablement les données étiquetées. Soit $\mathcal{X} = \mathcal{A} \cup \mathcal{N}$ un jeu de données, où $\mathcal{A}$ désigne le sous-ensemble d'anomalies étiquetées et $\mathcal{N}$ désigne le sous-ensemble non étiqueté. Le modèle génère des paires d'observations échantillonnées aléatoirement à partir de l'ensemble $\mathcal{A}$ et de l'ensemble $\mathcal{N}$. Ces paires sont classées en trois types :

1. **Paires anomalie-anomalie (AA)**. Les deux instances sont des anomalies étiquetées.
2. **Paires anomalie-non étiqueté (AN)**. Une instance est une anomalie étiquetée et l'autre provient de l'ensemble non étiqueté.
3. **Paires non étiqueté-non étiqueté (NN)**. Les deux instances proviennent de l'ensemble non étiqueté.

Chaque paire reçoit une étiquette ordinale synthétique en fonction de sa composition : $y_{AA} > y_{AN} > y_{NN}$. Cette classification guide le processus d'apprentissage en fournissant une hiérarchie de probabilités d'anomalie. Avec ces paires d'instances et leurs étiquettes correspondantes, PReNET crée un grand ensemble de données entièrement étiquetées $\mathcal{P} = \{(x_i, x_j, y_{x_i,x_j}) | x_i, x_j \in \mathcal{X}, y_{x_i,x_j} \in \mathbb{N}\}$, ce qui augmente significativement la quantité de données d'entraînement disponibles. En outre, $\mathcal{P}$ fournit une diversité d'exemples représentant différents écarts entre normalité et anormalité, ce qui aide le modèle à mieux distinguer les anomalies des instances normales. Cette diversité permet au modèle de s'adapter à des variations subtiles et complexes dans les données, rendant les représentations apprises plus robustes pour la détection d'anomalies.

Apprentissage de scores d'anomalies de bout en bout Le noyeau de PReNET réside dans son mécanisme d'apprentissage de bout en bout, où le modèle apprend directement à prédire les scores d'anomalies en comparant les échantillons par paire à l'aide d'un réseau de score d'anomalie à deux flux. Le réseau $\phi((\cdot,\cdot);\Theta) : \mathcal{P} \mapsto \mathbb{R}$ se compose de deux composants principaux :

1. **Apprenant de caractéristiques** (ψ). Cette partie du réseau transforme chaque instance de données en un espace de représentation intermédiaire $\mathcal{Z}$:

$$\psi(\cdot;\Theta_r) : \mathcal{X} \mapsto \mathcal{Z}$$

2. **Apprenant de score d'anomalie** (η). Ce composant prend la paire de représentations intermédiaires et calcule un score d'anomalie :

$$\eta(\cdot;\Theta_s) : (\mathcal{Z}, \mathcal{Z}) \mapsto \mathbb{R}$$

Ainsi, le modèle est formellement défini comme suit :

$$\phi((x_i, x_j);\Theta) = \eta((\psi(x_i;\Theta_r), \psi(x_j;\Theta_r));\Theta_s)$$

où $\Theta = \{\Theta_r, \Theta_s\}$ représente les paramètres de l'apprenant de caractéristiques et de l'apprenant de score d'anomalie.

Régression ordinale L'objectif d'apprentissage est de minimiser la perte de régression ordinale, qui est la différence absolue entre les scores prédits et les étiquettes de classe ordinale :

$$L(\phi((x_i, x_j); \Theta), y_{(x_i,x_j)}) = \left| y_{(x_i,x_j)} - \phi((x_i, x_j); \Theta) \right|$$

Cette fonction de perte garantit que le modèle attribue des scores plus élevés aux paires contenant des anomalies, différenciant ainsi efficacement les instances normales des instances anormales.

Calcul du score d'anomalie Lors de la phase de détection, PReNET calcule le score d'anomalie pour une nouvelle observation x_k en faisant la moyenne des scores obtenus par l'appariement avec des anomalies étiquetées et des instances non étiquetées :

$$s_{x_k} = \frac{1}{2E} \left(\sum_{i=1}^{E} \phi((a_i, x_k); \Theta^*) + \sum_{j=1}^{E} \phi((x_k, u_j); \Theta^*) \right)$$

Ici, Θ^* désigne les paramètres du modèle entraîné, et E est la taille de l'ensemble. Cette approche par ensemble stabilise le score d'anomalie, en tirant parti de la loi des grands nombres pour réduire la variance.

PReNET présente plusieurs avantages qui en font un choix incontournable pour la détection d'anomalies, en particulier dans les données textuelles. Tout d'abord, il dépasse les méthodes non supervisées traditionnelles en termes de performance en s'appuyant sur un nombre limité d'échantillons étiquetés. Ainsi, PReNET est très économe en termes de ressources et offre une meilleure exactitude. Deuxièmement, le mécanisme d'apprentissage par paire combiné à la régression ordinale permet à PReNET de bien se généraliser aux anomalies non vues, améliorant ainsi sa robustesse et sa fiabilité. Enfin, PReNET est particulièrement adapté à la détection d'anomalies textuelles grâce à sa capacité à gérer des données complexes. En s'appuyant sur l'information relationnelle entre les paires d'exemples, il capture des variations fines et des anomalies cachées dans des corpus textuels complexes, ce qui en fait un outil puissant pour la détection d'anomalies dans les textes.

3.4 Synthèse

Dans ce chapitre, nous avons présenté la méthodologie pour la détection d'anomalies textuelles, en nous concentrant sur deux composantes principales : les techniques de représentation de texte et les algorithmes de détection d'anomalies. Ce double focus nous a permis de relever les défis uniques posés par les données textuelles, qui sont intrinsèquement non structurées et sémantiquement complexes par rapport à d'autres types de données, telles que les données numériques ou tabulaires.

Nous avons commencé par discuter des techniques de représentation de texte qui convertissent les textes bruts en formats numériques appropriés pour l'apprentissage automatique. Les méthodes traditionnelles comme TF-IDF, bien qu'efficaces en termes de calcul, ne parviennent souvent pas à capturer les nuances contextuelles du texte. En revanche, des méthodes avancées comme Sentence-BERT, qui utilisent l'apprentissage profond pour générer des représentations contextuelles, se sont révélées plus efficaces pour saisir les relations sémantiques, ce qui les rend particulièrement utiles pour la détection d'anomalies textuelles.

Algo	Référence	Paradigme	Score	Arch.
ABOD	Kriegel et al. 2008	unsup.	prox. & prob.	ML
ALAD	Zenati et al. 2018	semi-sup. NO unsup.	reconst.	DL
COPOD	Li et al. 2020b	unsup.	stat.	ML
DevNet	Pang et al. 2019	semi-sup. PU	*end-to-end*	DL
ECOD	Li et al. 2022c	unsup.	stat.	ML
PReNet	Pang et al. 2020	semi-sup. PU	prox. & *end-to-end*	DL
XGBOD	Zhao and Hryniewicki 2018	weakly-sup. sup.	ensemble	ML

TABLE 3.1 – Bilan des algorithmes sélectionnés pour la détection d'anomalies textuelles. **Paradigme** : non supervisé (unsup.), semi-supervisé avec uniquement des données normales (semi-sup. NO), semi-supervisé avec données positives et non étiquetées (semi-sup. PU), faiblement supervisé (weakly-sup.), entièrement supervisé (sup.). **Mécanisme de score** : basé sur la proximité (prox.), la probabilité (prob.), les statistiques (stat.), la reconstruction (reconst.), l'apprentissage de score de bout en bout (*end-to-end*), l'apprentissage ensembliste (ensemble). **Architecture** : modèle profond (*deep learning*, DL), modèle peu profond (*machine learning*, ML). Ce tableau résume chaque algorithme par l'article original, le paradigme d'apprentissage, le mécanisme de score et l'architecture.

La seconde partie du chapitre s'est concentrée sur divers algorithmes de détection d'anomalies qui opèrent sur ces représentations de texte. Comme résumé dans le Tableau 3.1, nous avons sélectionné des algorithmes reflétant les tendances récentes de la recherche :

1. **Améliorations des méthodes traditionnelles** : Des algorithmes tels que COPOD et ECOD améliorent les méthodes statistiques traditionnelles pour mieux gérer les données à haute dimension.
2. **Mécanismes innovants de calcul de scores** : Des modèles comme DevNet et PReNET introduisent de nouveaux cadres d'apprentissage de bout en bout qui apprennent et optimisent les scores d'anomalie directement à partir des données d'entrée.
3. **Approches d'apprentissage ensembliste** : XGBOD intègre l'apprentissage ensembliste en combinant plusieurs détecteurs non supervisés avec un classificateur supervisé pour améliorer l'exactitude de détection.
4. **Approches hybrides de calcul de scores** : Des modèles comme ABOD combinent différents mécanismes de calcul de scores, tels que des approches basées sur la proximité et la probabilité, pour améliorer les performances.
5. **Intégration de la supervision** : Des modèles semi-supervisés et faiblement supervisés, incluant DevNet et XGBOD, exploitent un nombre limité de données étiquetées pour guider le processus de détection d'anomalies, améliorant ainsi les performances des modèles.
6. **Intégration de l'apprentissage profond** : L'utilisation de techniques d'apprentissage profond, exemplifiées par des modèles comme ALAD, permet de détecter des patterns complexes dans les données à haute dimension que les méthodes traditionnelles pourraient négliger.

Ce chapitre établit un cadre global pour la détection d'anomalies textuelles, combinant des techniques avancées de représentation de texte avec des algorithmes de

détection d'anomalies à l'état de l'art. Le chapitre suivant explorera les jeux de données utilisés dans notre étude, en détaillant les caractéristiques des corpus et les contextes dans lesquels ces méthodologies sont appliquées. Cette progression prépare le terrain pour une validation empirique et des analyses approfondies des techniques de détection d'anomalies textuelles.

CHAPITRE 4

DONNÉES

Sommaire

4.1 Introduction

Le développement de corpus efficaces est essentiel pour faire avancer la recherche en détection d'anomalies textuelles, mais cette entreprise rencontre des défis considérables. Comme discuté dans §2.3, la création de corpus adaptés à cette tâche est entravée par plusieurs obstacles. Tout d'abord, les anomalies sont intrinsèquement rares et imprévisibles, ce qui complique l'acquisition de suffisamment d'échantillons positifs (anomalies) pour construire un corpus représentatif pour la plupart des applications réelles. De plus, la définition de ce qui constitue une anomalie dépend fortement du contexte ; un texte considéré comme anormal dans un scénario peut être normal dans un autre, ce qui rend la création d'un corpus universellement applicable difficile. En outre, le processus d'annotation de tels corpus implique souvent un haut degré de subjectivité, en particulier dans les scénarios complexes où le contexte et les nuances du texte influencent grandement la perception et l'interprétation des anomalies. Cette subjectivité rend l'annotation des corpus coûteuse en temps et en ressources.

Malgré ces défis, des recherches académiques antérieures ont proposé plusieurs méthodologies pour atténuer ces problèmes (voir §2.3 pour une discussion détaillée). Ces stratégies peuvent être catégorisées en trois groupes principaux : 1) la synthèse, qui introduit des anomalies artificiellement créées dans un corpus existants dont le

textes sont considéré comme normaux ; 2) la fusion, qui combine diverses sources textuelles pour compiler un jeux de données avec des échantillons « normaux » provenant d'un corpus et des anomalies d'un autre ; et 3) l'adaptation consiste à transformer des corpus existants, initialement destinés à d'autres tâches, pour les utiliser dans la détection d'anomalies.

Dans notre travail, nous avons utilisé cette troisième stratégie d'adaptation des corpus en exploitant 14 jeux de données conçus à l'origine pour trois tâches différentes : **classification thématique**, **analyse de sentiments** et **détection de discours haineux**. À partir de ces 14 jeux de données, nous avons créé 17 corpus spécifiquement adaptés pour la détection d'anomalies textuelles. Ce chapitre examine en détail ces corpus et la logique de leur adaptation.

4.2 Sélection de jeux de données

Lors de la sélection des jeux de données, nous avons pris en compte plusieurs considérations essentielles. Tout d'abord, les anomalies textuelles se manifestent sous diverses formes et à différents niveaux linguistiques, allant des erreurs d'orthographe et grammaticales aux pseudépigraphes dans un anthologie. Pour faciliter une évaluation globale des mécanismes de détection d'anomalies, il convient d'inclure un large éventail de types d'anomalies. Étant donnée le contexte industriel de notre travail, nous avons spécifiquement ciblé trois types d'anomalies textuelles pertinentes pour la veille : les thèmes nouveaux (ou anormaux) cruciaux dans l'intelligence stratégique, les sentiments négatifs liés à la surveillance de la réputation en ligne, les discours de haine pertinents pour la cybersécurité.

De plus, alors que la majorité des recherches existantes en détection d'anomalies se sont limitées à l'anglais, notre travail étend la portée linguistique en incorporant des jeux de données en anglais, français et chinois. Cette approche multilingue élargit non seulement l'applicabilité de nos résultats mais approfondit également notre compréhension des anomalies textuelles dans différents contextes culturels et linguistiques.

En outre, pour couvrir autant de genres de textes que possible, nos données sélectionnées proviennent de sources diverses, y compris des agences de presse comme ABC News et Reuters, des plateformes de réseaux sociaux incluant Twitter et Weibo, et divers sites web comme Amazon et IMDB. Cette source diversifiée garantit que notre analyse est robuste et reflète les complexités des applications du monde réel.

Dans la discussion suivante, nous examinerons chaque jeu de données en détail, en décrivant leurs caractéristiques spécifiques et les raisons de leur sélection pour cette recherche.

4.2.1 Classification thématique : 20NG, AGNews et Reuters

La détection d'anomalies peut souvent être considérée comme un cas spécifique dans le cadre de la classification à classe unique. Les jeux de données de classification s'avèrent ainsi idéaux pour les tâches de détection d'anomalies grâce à leur facilité de s'adapter aux scénarios de classification à classe unique [Ruff et al., 2019; Manolache et al., 2021]. Dans des recherches récentes, divers jeux de données de classification thématique, notamment AGNews, 20NewsGroups et Reuters, ont été adaptés et utili-

sés comme benchmarks pour la détection d'anomalies textuelles [Barrett et al., 2019; Han et al., 2022; Pantin et al., 2022; Bejan et al., 2023].

AGNews Le jeu de données *AG News Topic Classification* (AGNews) [Zhang et al., 2015] est un sous-ensemble du plus large *AG's Corpus of News Articles* [1] qui est constitué des données collectées par le moteur de recherche académique ComeToMyHead [2] depuis juillet 2004. Le jeu de données AG original comprend plus d'un million d'articles de presse provenant de plus de 2 000 sources de nouvelles sur une période d'un an. À partir de cette vaste collection, le sous-ensemble AGNews a été créé en sélectionnant 496 835 articles classés sous quatre catégories : `Business`, `Science`, `Sports`, et `World`, chacune contenant 30 000 échantillons d'entraînement et 1,900 échantillons de test. Pour la recherche en classification de textes, seuls les titres et descriptions des articles sont conservés dans ce sous-ensemble. AGNews a été largement utilisé dans la recherche en classification de textes ainsi que dans les travaux de *benckmarking* en détection d'anomalies [Han et al., 2022; Pantin et al., 2022; Bejan et al., 2023].

20NG Le jeu de données *20 Newsgroups* (20NG) [3] est une vaste collection d'environ 20 000 documents de groupes de discussion, répartis uniformément entre 20 groupes de discussion distincts. Chaque groupe représente un sujet différent, allant des systèmes informatiques aux sports et à la politique. Le jeu de données offre un spectre diversifié de sujets, certains groupes étant étroitement liés, comme `comp.sys.ibm.pc.hardware` et `comp.sys.mac.hardware`, tandis que d'autres, tels que `talk.politics.mideast` et `misc.forsale`, sont nettement sans rapport. Cette granularité fine, en fait une ressource idéale pour explorer la détection d'anomalies dans des scénarios de paires *inliner-outlier* subtilement nuancées.

Reuters Le jeu de données *Reuters-21578* (Reuters) est une collection d'articles du service de nouvelles financières Reuters, compilée en 1987. Collectés et annotés par le *Carnegie Group, Inc.* et *Reuters, Ltd.*, ils étaient initialement conçus pour soutenir le développement du système de classification de textes `CONSTRUE`. Avec 10 369 documents couvrant une variété de sujets en finance, commerce et économie, Reuters a été l'un des jeux de données les plus influents et les plus utilisés dans la recherche en classification de textes. Le jeu de données est connu pour son approche d'annotation multi-étiquettes, une caractéristique peu commune dans des jeux de données de classification thématique qui attribuent typiquement une seule étiquette à chaque document. Les étiquettes couvrent plus de 100 sujets, englobant des domaines économiques et industriels tels que l'agriculture (`"rice"`, `"tea"`, `"soybean"`), la macroéconomie (`"income"`, `"gnp"`, `"money-supply"`) et diverses industries (`"gas"`, `"oil"`, `"steel"`). Cette annotation multi-classe rend Reuters extrêmement utile pour développer et tester des algorithmes capables de gérer des jeux de données complexes du monde réel. Bien que Reuters ait été supplanté par les collections plus modernes comme RCV1 [4], son impact profond sur le domaine de la classification de textes demeure. Il sert toujours de benchmark essentielle pour de nouveaux algorithmes et

1. http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html
2. http://newsengine.di.unipi.it/
3. http://qwone.com/~jason/20Newsgroups/
4. https://trec.nist.gov/data/reuters/reuters.html

méthodologies dans l'analyse de données textuelles, en particulier dans la détection d'anomalies [Han et al., 2022; Bejan et al., 2023].

4.2.2 Classification thématique : Covid-News (fr) et TTNews (cn)

TTNews Le jeu de données *TouTiao Text Classification for News Titles* (TTNews) [Xu et al., 2020] comprend une collection de 73 360 titres de nouvelles en chinois provenant de la plateforme TouTiao[5] (头条, « la une ») pour une période allant jusqu'à mai 2018. Les titres sont classés dans 15 catégories distinctes, telles que la finance, la technologie, la culture et le militaire. Ce jeu de données est conçu pour la classification de textes courts et l'identification des thématiques des nouvelles. TTNews fait partie des efforts du projet CLUE Benchmark, qui vise à créer un système de référence robuste pour la compréhension de la langue chinoise, similaire aux benchmarks GLUE [Wang et al., 2018a] et SuperGLUE [Wang et al., 2019a] en anglais [Xu et al., 2020]. Pour rendre le jeu de données plus discriminant, les auteurs ont utilisé une validation croisée à quatre plis pour filtrer certains des exemples les plus faciles. Un modèle ALBERT-tiny [Lan et al., 2020] est ajusté sur trois plis du jeu de données, puis utilisé pour sélectionner et filtrer les exemples moins complexes à partir du pli restant, garantissant ainsi un haut niveau de complexité pour les tâches de classification de texte.

COVIDNews Le jeu de données *COVID-19 French News* (COVIDNews) Cortal [2022] comprend une collection diversifiée de plus de 40 000 articles de presse liés à la COVID-19, recueillis à partir de plus de 50 sources d'actualités francophones en ligne. Collecté et compilé à l'aide de l'outil d'extraction d'informations news-please, ce jeu de données est élaboré pour la classification de textes au niveau des sujets, avec des étiquettes couvrant 12 catégories, telles la que santé, la société et la finance.

4.2.3 Classification thématique au niveau des événements : TDT2

Le jeu données *TDT2 Multilanguage Text V4.0* (TDT2)[6] [Cieri et al., 1999b; Fiscus et al., 1999] a été développé dans le cadre du programme *Topic Detection and Tracking* (TDT) financé par la DARPA, qui vise à développer des technologies pour identifier des matériaux liés à des thèmes dans des flux de données tels que les fils de presse et les émissions radiodiffusées. Créé pour soutenir la deuxième phase du programme TDT, le corpus TDT2 se concentre sur trois tâches clés : la segmentation (identifier des sections homogènes au niveau thématique), la détection (identifier de nouveaux événements) et le suivi (surveiller la récurrence des événements).

La tâche de détection, mieux connu sous le nom de *First Story Detection* (FSD) [Allan et al., 2000], est largement reconnue comme la première tentative systématique pour relever le défi de la détection de nouveauté dans les données textuelles [Tsai, 2010; Ghosal et al., 2022]. Le FSD consiste à identifier le premier reportage sur un événement. L' « événement » a été initialement défini comme « un incident spécifique qui se produit à un moment et en un lieu déterminés », puis étendu à « un incident

5. TouTiao, une application développée par ByteDance, est l'une des principales plateformes d'actualités et d'informations en Chine, offrant aux utilisateurs un flux personnalisé de titres de nouvelles. Cette plateforme est largement utilisée pour sa capacité à agréger et distribuer du contenu à grande échelle.

6. https://catalog.ldc.upenn.edu/LDC2001T57

séminal ainsi que toutes les activités directement liées », pour s'aligner sur la notion de *topic* » dans la recherche TDT [Allan et al., 2000; Harman et al., 2002]. Par exemple, la notion d'événement englobe non seulement des incidents majeurs comme les « attentats de Paris en novembre 2015 », mais aussi des activités subséquentes liées, y compris les réactions du gouvernement et les procédures légales suivantes.

TDT2 est un jeu de données multilingue qui comprend des données collectées quotidiennement durant six mois (janvier - juin 1998), à partir de sources en anglais américain et en chinois mandarin. Les sources de données en anglais incluent *Associated Press*, *New York Times*,*Public Radio International*, *Voice of America*, *ABC* et *CNN*, avec des transcriptions automatiques et manuelles disponibles pour les nouvelles radiodiffusées. Le jeu de données contient plus de 10 000 textes répartis sur 216 événements. Pour les 100 premiers événements, il fournit une documentation détaillée, incluant des informations sur le lieu, la date et une brève description de chaque événement.

N°	Texte	Étiquette
1	*The Indonesian President Suharto said today that Indonesians are going to have to make "painful sacrifices" to survive the economic crisis which is very rapidly plunging Indonesia closer to chaos ...*	`Asian Economic Crisis`
2	*good evening. defense secretary william cohen said today that a military strike against iraq would be "substantial in impact." but he stressed that the strike would not be able to remove saddam hussein from power ...*	`Current Conflict with Iraq`
3	*Israeli and Palestinian security forces report they have carried out separate raids on hide-outs of the militant Palestinian group Hamas where terrorist attacks were being planned ...*	`Israeli Palestinian Raids`
4	*After months of deadlock, the Israeli and Palestinian leaders have agreed to attend Mideast peace talks in London next month hosted by the United States and Britain ...*	`Israeli-Palestinian Talks (London)`
5	*Still overseas, in Indonesia today the latest signs that all is not well between the government and the governed. President Sue harto had no hesitation putting his riot police on the streets at the slightest signs of demonstration against his government ...*	`Anti-Suharto Violence`
6	*A China Airlines A-300 jetliner returning from the Indonesian island of Bali with 197 passengers and crew crashed and burst into flame Monday night just short of Taipei's Chiang Kai-shek airport ...*	`China Airlines Crash`

TABLE 4.1 – Exemples tirés du jeu de données TDT2 anglais, illustrant divers événements et leurs étiquettes associées.

TDT2 est choisi comme notre jeu de données principal tout au long de notre re-

cherche en raison de son annotation détaillée. Dans le contexte de détection d'anoamlies, les jeux de données de classification de texte traditionnellement utilisés, comme 20NG et AGNews, sont annotés à un niveau grossier et couvrent des catégories génériques telles que « sports », « politique », « affaires » et « science ». En revanche, TDT2 (voir le Tableau 4.1) se concentre sur des événements spécifiques tels que la « crise économique asiatique » ou les « jeux olympiques d'hiver de 1998 », ce qui facilite le test des méthodes sur des scénarios moins distinguables à travers des paires *inline-outlier* sémantiquement nuancées, telles que « raids israélo-palestiniens » contre « pourparlers de paix israélo-palestiniens à Londres ». Cette granularité fine est aussi essentielle pour simuler des scénarios réels où les anomalies doivent être détectées dans des contextes de sujets étroitement liés.

De plus, le développement de TDT2 sous la supervision rigoureuse de la DARPA et du NIST, ainsi que son enregistrement auprès de la LDC, garantit des annotations de haute qualité et bien documentées, le distinguant des jeux de données avec des processus de collecte et d'annotation moins transparents.

4.2.4 Analyse de sentiments : IMDB, Amazon, Yelp

Les jeux de données d'analyse de sentiments sont également utilisés dans la détection d'anomalies, notamment dans des scénarios où comprendre les changements de sentiments est crucial, comme le développement de systèmes de surveillance de la réputation en ligne et de plateformes de service client. Dans ces contextes, les sentiments négatifs sont souvent traités comme des anomalies qui déclenchent des alertes, ce qui signale des problèmes potentiels nécessitant une attention. Dans notre recherche, nous avons choisi trois ensembles de données d'analyse de sentiments importants : IMDB, Amazon et Yelp Review, chacun offrant des perspectives uniques sur différents scénarios.

IMDB *Internet Movie Database* est une plateforme en ligne qui permet aux utilisateurs de noter les films sur une échelle de 1 à 10 et de fournir des commentaires. Le *Large Movie Review Dataset* (IMDB) [Maas et al., 2011] comprend 50 000 commentaires de films collectés sur la plateforme et est spécifiquement conçus pour la classification binaire des sentiments. Chaque commentaire dans le jeu de données est étiqueté comme positif ou négatif, selon le sentiment exprimé par le commentaire. Les commentaires sont catégorisés comme négatifs s'ils obtiennent une note de 4 sur 10 ou moins, et comme positifs si la note est de 7 sur 10 ou plus, excluant délibérément les commentaires neutres pour accentuer le contraste de polarité. Ce jeu de données est parfaitement équilibré avec un nombre égal de commentaires positifs et négatifs, et est divisé également en sous-ensembles d'entraînement et de test. Le jeu de données IMDB dépasse significativement les benchmarks antérieurs tels que le *Polarity Dataset v2.0* [7] en termes d'échelle et de portée, ce qui en fait l'un des jeux de données les plus utilisés dans la classification binaire des sentiments.

Amazon Le jeu de données *Multilingual Amazon Reviews* (Amazon) [Keung et al., 2020] est une vaste collection de commentaires sur la plateforme Amazon, conçue pour les tâches de classification de texte. Ce corpus contient des commentaires dans six langues : anglais, japonais, allemand, français, chinois et espagnol, recueillies de

7. https://www.cs.cornell.edu/people/pabo/movie-review-data/.

novembre 2015 à novembre 2019. Chaque entrée dans le jeu de données comprend plusieurs composants : le texte du commentaire, le titre, la note en étoiles, l'identifiant du client, l'identifiant du produit et la catégorie générale de produit telle que `livres` ou `appareils`. Pour chaque langue, il y a 200 000, 5 000 et 5 000 commentaires dans les sous-ensembles d'entraînement, de développement et de test respectivement. Le corpus maintient un équilibre à travers le système de notation à cinq étoiles, chaque notation représentant 20% des commentaires. Les commentaires sont standardisées en longueur, avec une coupure à 2 000 caractères et un minimum de 20 caractères, pour garantir la cohérence de l'ensemble de données. Ce corpus a été largement utilisé dans une variété d'applications de fouille de textes, notamment dans la détection d'anomalies textuelles. Dans de tels contextes, il est souvent transformé en corpus de détection de polarité de sentiments, où le système d'évaluation à 5 étoiles est adapté en un cadre de polarité binaire.

Yelp Yelp.com est une plateforme en ligne où les utilisateurs peuvent partager leurs expériences en écrivant des commentaires et en notant diverses entreprises locales. En 2015, Yelp Inc. a lancé le *Yelp Open Dataset* dans le cadre de son projet *Dataset Challenge*. Ce vaste jeu de données est initialement compilé pour fournir des informations complètes sur le comportement des consommateurs et les caractéristiques des entreprises. Il comprend 1,6 million de commentaires et 500 000 conseils émanant de 366 000 utilisateurs pour 61 000 entreprises. Outre les commentaires et les conseils, l'ensemble de données comprend des données supplémentaires telles que les heures d'ouverture des entreprises, la disponibilité des parkings et les visites des utilisateurs. Le *Yelp Review Dataset* (Yelp Review) est un sous-ensemble du *Yelp Open Dataset* qui est centré sur les commentaires des utilisateurs. Ce sous-ensemble contenant deux champs : le texte (commentaire) et l'étiquette (note allant de 1 à 5).

4.2.5 Détection de discours haineux : OLID, COLDataset, MLMA

Les jeux de données de détection de discours haineux ou de langage offensant constituent une autre ressource précieuse pour la recherche en détection d'anomalies. La détection des anomalies est particulièrement adaptée à l'identification de ces formes de communication en raison de la subtilité et de la nature contextuelle de leur interprétation [Wahl, 2021]. Contrairement aux méthodes traditionnelles basées sur des approches de classification, qui reposent sur des données étiquetées et peuvent avoir des difficultés à capturer les complexités évolutives de la langue, la détection d'anomalies excelle en se concentrant sur les points de données aberrants, c'est-à-dire les éléments qui s'écartent de manière substantielle de la norme établie. En apprenant des patterns du langage non offensant et en repérant les écarts significatifs, la détection d'anomalies s'adapte de manière dynamique. Cela permet de reconnaître efficacement diverses formes de langage nuisible sans s'appuyer sur des définitions rigides, répondant ainsi à la nature subjective de ce qui constitue un contenu offensant. Dans notre recherche, nous avons intégré trois corpus distincts de discours de haine et de langage offensant : OLID, COLDataset et MLMA.

OLID *Offensive Language Identification Dataset* (OLID) [Zampieri et al., 2019] est un jeu de données hiérarchique conçu pour l'identification et la catégorisation du langage offensant sur les réseaux sociaux, en particulier Twitter. Composé de 14 100 tweets en anglais, dont 13 240 pour l'entraînement et 860 pour les tests, OLID uti-

lise un schéma d'annotation hiérarchique à trois niveaux pour une analyse nuancée : 1) Détection de langage offensant (`offensant` et `non offensant`), 2) Catégorisation du langage offensant (`insultes ciblées` et `contenu non ciblé`), et 3) Identification de la cible du langage offensant (ciblant un `individu`, un `groupe` ou une `autre entité`). La création d'OLID est réalisée à l'aide de l'API de Twitter pour récupérer sélectivement des tweets basés sur des mots-clés et des phrases couramment trouvés dans les messages offensants, avec un accent sur les termes politiquement chargés tels que « *@BreitBartNews* », « *MAGA* », « *antifa* » et « *liberals* », qui présentent souvent une incitation plus forte de langage offensant.

COLDataset Le jeu de données *Chinese Offensive Language Dataset* (COLDataset) [Deng et al., 2022] est conçu pour l'analyse et la détection du langage offensant sur les plateformes de réseaux sociaux chinois. Les données ont été recueillies sur les plateformes chinoises telles que Zhihu(知乎) et Weibo(微博). Comme les posts sont majoritairement non-offensants, la collection de données a été réalisée à l'aide de requêtes par mots-clés pour capturer du contenu riche en langage offensant. Ce jeu de données comprend 37 480 phrases annotées, ciblant des problématiques de préjugés de race, de sexe et de région. Le jeu d'entraînement inclut 32 157 phrases semi-automatiquement étiquetées pour leur caractère offensant, en utilisant un système de modèle en boucle (*model-in-the-loop*) pour améliorer l'efficacité. Le jeu de test, composé de 5 323 phrases, est annoté par des experts humains en quatre catégories : `attaque contre des individus`, `attaque contre des groupes`, `anti-biais`, et `autre message non-offensant`.

MLMA *Multilingual and Multi-Aspect Hate Speech Analysis* (MLMA) [Ousidhoum et al., 2019] est un jeu de données multilingue conçu pour la détection et l'analyse des discours de haine à travers différentes langues et cultures. Ce jeu de données comprend environ 13 000 tweets répartis entre l'anglais (5 647), le français (4 014) et l'arabe (3 353). Chaque tweet est annoté selon plusieurs dimensions clés : la directivité du discours de haine (`direct` ou `indirect`), le type d'hostilité (comme `offensant`, `irrespectueux` ou `haineux`), la cible spécifique (par exemple, l'`origine ethnique`, le `sexe`) et le groupe particulier ciblé (par exemple, les `immigrants`, les `groupes religieux`). La collecte des données a été organisée pour obtenir une représentation équilibrée des types de discours haineux les plus courants dans chaque langue, ce qui fait du MLMA une référence solide pour l'évaluation des systèmes de détection de discours haineux.

4.3 Création de corpus

L'adaptation est l'étape centrale de la préparation des corpus pour la détection d'anomalies. Dans cette section, nous décrivons les stratégies d'adaptation adoptées pour chaque corpus, en tenant compte des particularités de leurs textes et de leurs annotations. Avant d'entrer dans les détails de chaque corpus, nous introduisons d'abord les principes généraux de l'adaptation des jeux de données :

1. **Objectif général** : Les jeux de données sont initialement conçus pour trois tâches distinctes, à savoir la classification de thématiques, la détection de (polarité de) sentiments et la détection de discours haineux, avec des étiquettes multiclasse et une distribution équilibrée des données. Cependant, dans le cadre

de la détection d'anomalies, la composition des corpus doit refléter une classification binaire spécifique où la distribution des étiquettes est extrêmement déséquilibrée, marquée par une prédominance d'échantillons normaux et une rareté des anomalies. Par conséquent, l'adaptation implique principalement l'échantillonnage des données à partir des jeux de données originaux, le mapping des étiquettes multiclasse vers un espace d'étiquettes binaire, et l'ajustement de la proportion des étiquettes selon les taux d'anomalies définis.

2. **Ratio d'anomalie** (λ_a) : Dans les jeux de données du monde réel, les anomalies sont rares, ce qui conduit à des distributions de classes très déséquilibrées avec des ratios d'anomalie extrêmement faibles (souvent inférieurs à 2%). En revanche, dans la recherche académique, les jeux de données synthétiques ont souvent un taux d'anomalie contrôlé pour permettre une comparaison plus facile entre les études. Dans notre travail, nous avons normalisé le ratio d'anomalie λ_a à 10%, conformément aux pratiques courantes dans le domaine où les ratios sont généralement fixés à 1%, 5% ou 10% [Manolache et al., 2021; Pantin et al., 2022; Han et al., 2022; Ait-Saada and Nadif, 2023]. Cette décision s'aligne également sur les taux de contamination typiques utilisés dans les outils de détection d'anomalies populaires tels que Scikit-learn [Pedregosa et al., 2011], PyOD [Zhao et al., 2019] et DeepOD [Xu et al., 2023b].
3. **Échantillonnage** : Nous avons créé les corpus en effectuant un échantillonnage aléatoire stratifié à partir des jeux de données originaux, en respectant le ratio d'anomalie prédéfini. Dans les cas où les jeux de données étaient divisés en 2 ou 3 sous-ensembles, nous avons agrégé ces sous-ensembles et procédé à un échantillonnage à partir de l'ensemble de données.

Sur la base de ces principes, nous avons créé une série de corpus pour la détection d'anomalies textuelles à partir des jeux de données mentionnés ci-dessus (§4.2).

4.3.1 Anomalies au niveau des thématiques (événements)

TDT2 Le corpus TDT2, vu ses avantages décrits dans §4.2, est le corpus principal utilisé tout au long de cette thèse. Lors de la création de ce corpus, nous avons cherché à simuler un scénario réel pour la surveillance des médias, où un seul événement ciblé est considéré comme la norme, et tous les autres événements divergents sont traités comme des anomalies. L'objectif principal était d'établir un corpus dans lequel : 1) la classe normale est centrée sur un événement unique et bien documenté, et 2) la taille globale du corpus est aussi grande que possible, avec un taux d'anomalie fixé à 10%. Étant donnée qu'aucune des thématiques du sous-ensemble chinois ne contient plus de 100 documents, il est impossible de construire un corpus chinois qui réponde à ces critères. Par conséquent, notre recherche s'est concentrée sur le sous-ensemble anglais.

Le jeu de données anglais comprend environ 10 000 documents dispersés dans 216 thématiques au niveau de l'événement. La plupart de ces thématiques contiennent moins de 10 documents, mais trois d'entre elles comptent chacune plus de 1 000 documents. Nous avons sélectionné l'un de ces trois thématiques majeures pour former la base de notre classe normale, ce qui nous a permis de créer un corpus de détection d'anomalies de 1 000 documents.

Pour la classe normale, nous avons choisi la thématique « `Asian Economic Crisis` (thématique n°1) » comme thématique principale, à partir de laquelle nous

avons extrait 900 documents de manière aléatoire. En ce qui concerne la classe anormale, nous avons regroupé les documents des thématiques n°2 à n°40, en garantissant que chaque thématique au sein de cette classe anormale contribue à 2 ou 3 documents. Cette adaptation permet de créer un environnement de détection d'anomalies robuste, qui reproduit fidèlement la variabilité et le caractère inattendu des anomalies réelles dans le domaine de la surveillance des médias. Ainsi, nous avons constitué à partir du jeu de données original un corpus TDT2 standard.

TDT2-hard Outre le corpus standard, nous avons élaboré un corpus plus complexe, nommé TDT2-hard, conçu pour mieux refléter les défis des applications réelles. Dans de nombreuses applications pratiques, les anomalies ne sont pas toujours très éloignées du thème principal, ce qui rend leur détection difficile. Par exemple, des événements tels que des troubles sociaux ou des émeutes survenant pendant une crise économique peuvent sembler liés au thème central de la crise, mais ils représentent pourtant des anomalies importantes à détecter dans un contexte de veille. Le corpus TDT2-hard a été conçu pour simuler ces scénarios nuancés, où la distinction entre ce qui est « normal » et « anormal » est subtile et parfois floue. Cela reflète des situations réelles où des événements mineurs ou indirectement liés peuvent avoir un impact significatif et nécessitent une détection d'anomalies plus fine.

Pour cette variante plus difficile, nous avons conservé la « `Asian Economic Crisis` » comme thématique normale. Pour les thématiques anormales candidates, nous avons uniquement retenu les thématiques comportant plus de 50 documents. Après avoir transformé les textes via SBERT, nous avons calculé la distance cosinus entre les paires de textes, ce qui permet déterminer les distances moyennes entre les thématiques anormales candidates et le thématique normale. À l'issue de ce processus, la thématique n°76 « `Anti-Soharto Violence` », qui se concentre sur les troubles sociaux et politiques tels que les activités anti-gouvernementales durant la crise économique, s'est révélé comme la thématique la plus proche de la norme. Pour vérifier s'il s'agit effectivement du plus difficile à distinguer, nous avons utilisé les techniques KNN, LOF, et GMM pour calculer les scores d'anomalie pour chaque texte. Les analyses de la divergence de Kullback-Leibler et de la distance de Hellinger entre les distributions des scores d'anomalie ont confirmé que la thématique n°76 présente les valeurs les plus basses, indiquant une faible distinction par rapport à la norme. Ainsi, le TDT2-hard est composé de 900 textes la thématique n°1 (norme) et 100 textes la thématique n°76 (anomalie), ce qui offre un cadre robuste pour tester la détection d'anomalies dans des scénarios nuancés et complexes.

4.3.2 Anomalies au niveau des thématiques

Dans le cadre de la création des corpus pour la détection d'anomalies au niveau des thématiques à gros grains, les jeux de données AGNews, 20NG, Reuters, TTNews (chinois) et COVIDNews (français) ont été adaptés selon la méthodologie suivante :

1. Filtrer les classes trop petites (par exemple, moins de 1 000 documents) pour éviter qu'elles ne constituent la norme dans un contexte de détection d'anomalies.
2. Sélectionner n classes comme classes normales, en tenant compte de la nature de la classe et des caractéristiques spécifiques des textes.

3. Constituer n sous-ensembles correspondants, chacun avec une classe normale majoritaire et un taux de 10% d'anomalies sélectionnées aléatoirement parmi les autres classes.
4. Rassembler ces sous-ensembles pour former le corpus final.

Les corpus créés selon cette méthode se détaillent comme suit :

— **AGNews** : Quatre classes normales ont été retenues : `world`, `sports`, `business` et `science`. Chaque sous-ensemble contient 10 000 documents, avec 9 000 documents normaux et 1 000 anormaux, totalisant ainsi 40 000 documents pour l'ensemble du corpus.

— **20NG** : Uniquement les thématiques de niveau supérieur ont été conservés, en fusionnant des sous-thématiques tels que `comp.sys.ibm.pc.hardware` et `comp.sys.mac.hardware`. Les classes normales sélectionnées, `comp`, `rec`, `sci`, et `talk`, ont chacune constitué un sous-ensemble de 1 000 documents, comprenant 900 documents normaux et 100 anomalies, pour un total de 4 000 documents.

— **Reuters** : Après avoir éliminé les textes portant plusieurs étiquettes, les classes `acq` (*acquisitions*) et `earn` ont été désignées comme normales. Chaque sous-ensemble comprend 2 000 documents, dont 1 800 normaux et 200 anomaux, atteignant ainsi un total de 4 000 documents.

— **TTNews** : Les classes normales choisies sont `agriculture`, `technology`, `finance`, `sports` et `military`. Chaque sous-ensemble contient 2 000 documents, avec 1 800 documents normaux et 200 anomalies, totalisant ainsi 10 000 documents pour l'ensemble du corpus.

— **COVIDNews** : La classe normale choisie est la `santé`. Ce corpus ne comporte qu'une seule classe normale. Le corpus total comprend 5 000 documents, avec 4 500 documents normaux et 500 anomalies couvrant des sujets tels que le `sport` et la `technologie`.

En complément de ces corpus standard, deux corpus plus complexes ont été créés à partir de 20NG. Chaque corpus, nommé 20NG-hard-1 et 20NG-hard-2, contient 1 000 documents. Les sous-thématiques utilisées pour ces corpus sont les suivantes :

— **20NG-hard-1** : La classe normale est `comp.sys.ibm.pc.hardware` et la classe anormale est `comp.sys.mac.hardware`.

— **20NG-hard-2** : La classe normale est `talk.politics.misc` et la classe anormale est `talk.religion.misc`.

Ces corpus 'difficiles' sont conçus pour évaluer la robustesse des modèles face à des scénarios d'anomalies plus subtiles, offrant ainsi un défi significatif pour la détection d'anomalies.

4.3.3 Anomalies au niveau des sentiments

Pour la création des corpus destinés à la détection d'anomalies au niveau des sentiments, nous avons adapté les jeux de données Amazon (versions anglaise, chinoise, et française), IMDB et Yelp. La méthodologie d'adaptation dépend de la forme des étiquettes :

1. Pour les jeux de données où les sentiments sont étiquetés en termes de polarité, la classe des sentiments positifs a été considérée comme normale, et celle des sentiments négatifs comme anormale.

2. Dans les cas où une échelle de cinq points était utilisée, les textes évalués à 1 ou 2 points ont été classés comme anormaux (sentiments négatifs), et ceux évalués à 4 ou 5 points comme normaux (sentiments positifs).

Les corpus ainsi créés se présentent comme suit :

— **Amazon-(en/zh/fr)** : Trois corpus ont été créés, chacun pour une langue (anglais, chinois et français). Pour éviter les facteurs de confusion, chaque corpus utilise uniquement les commentaires d'une seule catégorie de produits : `apparel` pour les versions anglaise et chinoise, et `toy` pour la version française. Chaque corpus contient 5 000 documents, avec 4 500 documents normaux et 500 anomalies.
— **IMDB** : Le corpus contient 10 000 documents, avec 9 000 documents normaux et 1 000 anomalies.
— **Yelp** : Similaire à IMDB, ce corpus est également composé de 10 000 documents, dont 9 000 normaux et 1 000 anomalies.

En plus de ces corpus standard, nous avons créé des variantes pour les corpus Amazon, pour chaque langue, afin de tester la robustesse des modèles face à des scénarios de sentiments plus ou moins évidents :

— **Amazon-(en/zh/fr)-easy** : Dans cette variante plus facile, les commentaires avec 5 étoiles sont considérés comme normaux et ceux avec 1 étoile comme anormaux.
— **Amazon-(en/zh/fr)-hard** : Dans cette variante plus difficile, les commentaires avec 4 étoiles sont considérés comme normaux et ceux avec 2 étoiles comme anormaux.
— Chaque variante contient 2 000 documents.

4.3.4 Anomalies au niveau de l'usage du langage

La création des corpus pour la détection du discours haineux en tant qu'usage anormal du langage a impliqué l'adaptation des jeux de données OLID, COLDataset (chinois) et MLMA (français). Ces adaptations ont été orientées par la méthodologie suivante :

1. Dans les contextes de classification binaire, la classe positive (contenu haineux ou offensant) a été identifiée comme anormale, tandis que la classe négative (contenu non haineux ou non offensant) a été désignée comme normale. Cela s'aligne sur la pratique typique de détection d'anomalies où les expressions nuisibles sont considérées comme aberrantes.
2. Pour les scénarios de classification multi-classe, divers types de discours haineux ont été regroupés en une seule classe anormale, en contraste avec les textes non haineux regroupés dans la classe normale.

Détails des corpus adaptés pour cette étude :

— **OLID** : Ce corpus inclut uniquement les textes contenant plus de 10 tokens. La taille totale du corpus est de 5 000 documents.
— **COLDataset** : Seuls les textes issus du sous-ensemble d'entraînement, annotés comme offensants ou non-offensants, ont été utilisés. En outre, les textes comprenant moins de 10 caractères chinois ont été éliminés. Le corpus final contient 5 000 documents.

— **MLMA** : Dans ce corpus nous considérons des expressions offensives, irrespectueuses et des discours haineux comme anomalies, et les message neutres sans éléments nuisibles comme norme. Il comprend 900 documents.

4.4 Synthèse

Dans ce chapitre, nous avons présenté les données utilisées pour la détection d'anomalies textuelles, en détaillant les jeux de données sélectionnés et la méthode d'adaptation des corpus. La sélection des jeux de données a été guidée par la nécessité de couvrir un large éventail de types d'anomalies textuelles, notamment des anomalies thématiques, des anomalies sentimentales et des discours haineux. Les jeux de données incluent des collections en plusieurs langues (anglais, français, chinois) et proviennent de diverses sources, telles que les articles de presse, les réseaux sociaux et les plateformes de critiques, pour garantir une robustesse et une pertinence des résultats dans divers contextes.

Notre travail s'est concentré sur l'adaptation de 14 jeux de données existants afin de créer 17 corpus spécifiques à la détection d'anomalies textuelles. Cette adaptation a impliqué des modifications spécifiques pour introduire des anomalies contrôlées et ajuster les étiquettes des données, permettant ainsi de refléter les scénarios réels où les anomalies sont rares et souvent contextuelles. Les corpus ainsi créés sont conçus pour surmonter les défis associés à la rareté des anomalies et à la subjectivité de l'annotation des données.

Ce chapitre a ainsi établi une base solide pour les expérimentations futures, en fournissant des corpus diversifiés et adaptés qui seront utilisés pour évaluer l'efficacité des méthodes de détection d'anomalies dans le chapitre suivant consacré aux expériences.

CHAPITRE 5

EXPÉRIENCES

Sommaire

5.1 Introduction

Dans les chapitres précédents de cette partie, nous avons examiné en détail les fondements théoriques et les approches méthodologiques pertinentes pour la détection d'anomalies textuelles le cadre de fouille de données, ainsi que les jeux de données utilisés dans nos études. Ce chapitre est consacré à l'élaboration des procédures et configurations expérimentales, ainsi qu'à l'évaluation empirique des algorithmes de fouille de données adaptés.

Nos expériences sont conçues pour évaluer en profondeur l'efficacité et la robustesse des algorithmes de détection d'anomalies existants lorsqu'ils sont appliqués à des données textuelles. Cela permet d'explorer une solution potentielle pour la détection d'anomalies textuelles dans le cadre de fouille de données.

Ce chapitre se compose de deux sections :

1. **Configuration expérimentale** : Dans cette section, nous examinerons en détail la configuration expérimentale de nos études. Nous commençons par les processus de préparation des données, plus précisément le partitionnement du

corpus et la transformation des étiquettes, nécessaires pour les approches semi-supervisées et faiblement supervisées. Nous décrivons ensuite l'exécution des expériences, y compris les détails des essais indépendants réalisés et les pratiques concernant le paramétrage des hyperparamètres dans ce domaine. En fin, nous présenterons les métriques d'évaluation choisies dans notre étude, avec des justifications pour chaque choix.

2. **Résultats et discussion** : Cette section est précisément alignée sur les problématiques et les questions de recherche posées à l'Introduction de la deuxième partie. Elle présente une discussion approfondie apportant nos réponses à ces questions.

5.2 Configuration expérimentale

5.2.1 Répartition des données et validation croisée

Dans cette étude, nous avons élaboré 17 corpus en adaptant 14 jeux de données pour couvrir trois types d'anomalies : thématiques anormales, sentiments déviants et discours de haine. Pour assurer la robustesse et la fiabilité de nos résultats, nous avons employé une méthodologie de validation croisée à cinq plis.

Il s'agit de diviser aléatoirement les données en cinq plis. À chacune des cinq itérations, un pli est utilisé comme ensemble de test, tandis que les quatre plis restants servent de données d'entraînement. Cette procédure assure que chaque instance du jeu de données est utilisée à la fois pour l'entraînement et le test, offrant une évaluation complète des performances du modèle. Pour obtenir la prédiction finale sur l'ensemble du corpus, nous avons combiné les résultats de chaque itération. De plus, nous avons utilisé un échantillonnage stratifié pour maintenir un ratio d'anomalies constant dans tous les plis, assurant ainsi que chaque pli reflète fidèlement la distribution globale du jeu de données.

Ce processus de validation croisée a été répété deux fois, et nous avons calculé la moyenne des résultats pour chaque métrique d'évaluation. Cette moyenne agrégée minimise l'impact potentiel des variations aléatoires, fournissant ainsi une évaluation plus stable et fiable des performances du modèle.

5.2.2 Transformation des étiquettes

Dans notre étude, nous commençons avec un corpus entièrement annoté où chaque entrée est représentée par un tuple (texte, étiquette), avec des étiquettes « 0 » pour les instances normales et « 1 » pour les anomalies. Ce schéma d'étiquetage est applicable aussi bien aux scénarios supervisés qu'aux scénarios non supervisés, car les algorithmes non supervisés ignorent les étiquettes.

Cependant, pour l'apprentissage semi-supervisé et faiblement supervisé, il est nécessaire de transformer une partie de ces étiquettes, soit en les masquant, soit en les modifiant. Dans ce contexte, certaines étiquettes sont converties en inconnu (représenté par « -1 » ou « `None` », selon l'implémentation utilisée). La proportion d'étiquettes conservées est définie par le taux d'annotation γ_l, qui varie selon différents paramètres : 1%, 5%, 10%, 20%, 30%, ..., 90%, 100%.

— **Apprentissage semi-supervisé NO (*Normal-Only*)**. Pour l'apprentissage semi-supervisé NO, toutes les étiquettes d'anomalies « 1 » sont converties en

inconnues. Cette approche est utilisée pour tester la capacité des algorithmes à identifier les anomalies lorsqu'aucune étiquette d'anomalie explicite n'est fournie pendant l'entraînement.

— **Apprentissage semi-supervisé PU (*Positive-Unlabeled*)**. Pour l'apprentissage semi-supervisé PU, une partie des étiquettes positives « 1 » est conservée, tandis que les autres étiquettes sont changées en inconnues.

— **Apprentissage faiblement supervisé**. Pour l'apprentissage faiblement supervisé, nous conservons seulement une partie des étiquettes (à la fois « 0 » et « 1 »), les autres étant transformées en inconnues [1]. Pour maintenir un taux d'anomalie constant, nous utilisons un échantillonnage stratifié, assurant que le sous-ensemble retenu représente fidèlement la distribution globale des anomalies.

5.2.3 Hyperparamètres

Dans de nombreuses tâches d'apprentissage automatique, il est courant de faire tourner un algorithme plusieurs fois avec différents hyperparamètres pour identifier les réglages qui optimisent la performance. Ce processus, connu sous le nom de recherche d'hyperparamètres, est efficace dans les tâches de classification et de régression standard. Cependant, une telle approche s'avère problématique pour la détection d'anomalies en raison des caractéristiques uniques des anomalies [Aggarwal, 2017a]. Les anomalies sont, par définition, des événements qui s'écartent de manière significative des patterns habituels et qui ne sont pas prévisibles. Par conséquent, les hyperparamètres optimisés sur les données d'apprentissage échouent souvent à se généraliser aux nouvelles données, où les anomalies peuvent se manifester de manière imprévue. Cela peut conduire à un surajustement aux données d'entraînement, ce qui réduit l'efficacité du modèle dans des scénarios réels.

Pour garantir une évaluation fiable et une comparaison équitable, nous adhérons à la pratique courante qui consiste à utiliser les hyperparamètres par défaut tels que prescrits dans les publications originales pour chaque algorithme, à l'exception unique du taux de contamination [Aggarwal, 2017a; Han et al., 2022]. Les algorithmes de détection d'anomalies attribuent généralement un score d'anomalie qui est converti en un résultat binaire basé sur un seuil défini par le taux de contamination. Étant donné que le taux d'anomalie λ_a dans nos jeux de données est systématiquement fixé à 10%, nous standardisons le taux de contamination à 10% pour tous les algorithmes. Cette approche uniforme facilite une évaluation standardisée et équitable.

En conclusion, en évitant le réglage des hyperparamètres et en maintenant un taux de contamination constant, nous préservons la simplicité et l'équité de notre processus d'évaluation. Cela assure une comparaison équitable et directe entre différents algorithmes de détection d'anomalies.

5.2.4 Algorithmes de référence

Dans nos expériences, en plus des algorithmes décrits dans le Chapitre 3, nous avons également examiné plusieurs méthodes traditionnelles comme référence. Sui-

1. Il convient de noter que lorsque le taux d'annotation $\gamma_l = 100\%$, la tâche devient effectivement une tâche de classification supervisée, avec toutefois un ensemble de données extrêmement déséquilibré en raison de la rareté des anomalies.

Algorithme	Nom	Citation	Paradigme	Score	Architecture
AutoEncoder	*Auto Encoder*		non supervisé	reconstruction	DL
DeepSVDD	*Deep One-Class Classifier with AutoEncoder*	Ruff et al. [2018]	non supervisé	domaine	DL
DIF	*Deep Isolation Forest*	Xu et al. [2023b]	non supervisé	ensemble	DL
HBOS	*Histogram-based Outlier Detection*	Goldstein and Dengel [2012]	non supervisé	statistique-probabilité	ML
IForest	*Isolation Forest*	Liu et al. [2008]	non supervisé	ensemble	ML
kNN	*k-Nearest Neighbors Detector*	Ramaswamy et al. [2000]	non supervisé	proximité	ML
LOF	*Local Outlier Factor*	Breunig et al. [2000]	non supervisé	proximité	ML
OCSVM	*One Class Support Vector Machine*	Schölkopf et al. [2001]	semi-sup. NO	domaine	ML
PCA	*Principal Component Analysis*	Shyu et al. [2003]	non supervisé	reconstruction	ML

TABLE 5.1 – Aperçu des algorithmes de détection d'anomalies de référence.

vant les mêmes principes indiqués dans la § 3.1, nous avons cherché à couvrir la plupart des mécanismes traditionnels de détection d'anomalies, telles que les méthodes à base de proximité et à base de reconstruction. Si le chapitre méthodologique introduit un modèle d'apprentissage profond, nous optons ici pour un modèle traditionnel d'apprentissage automatique, et vice versa. Ce principe nous a conduits à explorer neuf méthodes différentes, comme détaillé dans le Tableau 5.1.

5.2.5 Cadre de l'apprentissage inductif

Dans nos expériences, nous examinons des algorithmes conçus pour différents paradigmes d'apprentissage : non supervisé, semi-supervisé (PU et NO), et faiblement supervisé (annotation partielle).

Les algorithmes non supervisés, par leur nature même, ont tendance à fonctionner de manière transductive, en se concentrant sur la prédiction pour un ensemble spécifique d'instances fournies pendant l'entraînement. Ils ne sont pas nécessairement censés se généraliser bien à de nouveaux ensembles de données. Cependant, pour assurer une comparaison juste et fiable avec les algorithmes avec supervision, nos expériences sont uniformément menées dans un cadre inductif. Dans ce contexte, les algorithmes non supervisés apprennent des paramètres et des patterns normaux à partir des données d'entraînement. Ensuite, ces modèles sont sérialisés et appliqués pour détecter des anomalies lors de la phase de test. Cette adaptation permet aux algorithmes non supervisés d'être utilisés de manière similaire aux algorithmes supervisés, facilitant ainsi une comparaison équitable. En effet, la plupart des implémentations existantes des algorithmes de détection d'anomalies ont adapté les méthodes non supervisées pour fonctionner dans un cadre inductif.

5.2.6 Métriques d'évaluation

En considérant les diverses métriques d'évaluation discutées dans le Chapitre 1, nous nous priorisons l'utilisation de l'AUCROC comme notre métrique principale dans les expériences et l'analyse de cette partie pour les raisons suivantes :

— **Indépendance du seuil** : L'AUCROC offre une perspective globale sur la capacité d'un modèle à distinguer entre les instances normales et anormales sans se lier à un seuil spécifique. Cette caractéristique est cruciale dans des scénarios où la définition du seuil optimal n'est pas triviale et où des seuils mal ajustés peuvent entraîner des résultats trompeurs.

— **Interprétation probabiliste** : L'interprétation probabiliste de l'AUCROC est l'un de ses aspects les plus intéressants. L'AUCROC peut être interprétée comme la probabilité qu'un échantillon positif (anomalie) obtienne un score d'anomalie plus élevé qu'un échantillon négatif, lorsque les deux échantillons

sont choisis de manière aléatoire. Ainsi, un AUCROC de 0,90 signifie qu'il y a 90% de chances que le détecteur attribue un score plus élevé à une anomalie choisie aléatoirement qu'à une instance normale choisie aléatoirement.

L'AUCROC évalue la performance globale d'un modèle en prenant en compte à la fois le taux de vrais positifs et le taux de faux positifs. Elle offre une vision plus large de la capacité du modèle à distinguer entre les instances normales et anormales, mais peut parfois s'avérer moins sensible aux performances concernant la classe minoritaire. Dans les jeux de données fortement déséquilibrés, l'AUCROC pourrait présenter une vision trop optimiste de la performance du modèle puisqu'elle inclut le taux de vrais négatifs (spécificité), qui est généralement élevé. En revanche, l'AUCPR offre une évaluation plus réaliste de l'efficacité du modèle à identifier la classe minoritaire. Il est plus informatif de compléter l'AUCROC avec l'AUCPR dans les scénarios fortement déséquilibrés où le nombre de cas positifs est bien inférieur au nombre de cas négatifs. Ainsi, pour une évaluation complète et précise dans la détection d'anomalies, nous avons choisi d'adopter l'AUCPR comme métrique complémentaire.

5.3 Résultats et analyses

Dans cette section, nous présentons une analyse et une discussion détaillées des résultats dérivés d'une série d'expériences conçues pour faire progresser notre compréhension de la détection d'anomalies textuelles à l'aide de techniques de fouille de données [2].

Nous avons évalué 17 algorithmes distincts de détection d'anomalies, chacun basé sur divers cadres théoriques et paradigmes d'apprentissage. Nos expériences ont exploité une grande variété de corpus dans trois langues, en se concentrant sur trois types spécifiques d'anomalies textuelles : les anomalies thématiques, les sentiments déviants et les discours de haine. De plus, les données ont été représentées à l'aide de techniques traditionnelles et de méthodes modernes de plongements. Ces configurations ont permis une évaluation complète de l'efficacité des algorithmes de détection d'anomalies dans les données textuelles.

L'objectif principal de cette analyse est de répondre aux questions de recherche énoncées au début de la Partie II.

5.3.1 Paradigme d'apprentissage

Nous commençons par examiner l'impact des différents paradigmes d'apprentissage automatique sur l'efficacité de la détection d'anomalies dans les textes, en nous concentrant plus particulièrement sur l'utilisation de données partiellement étiquetées. Notre analyse porte sur trois paradigmes d'apprentissage distincts : les approches non supervisées, semi-supervisées (incluant *Normal-Only* ou NO, et *Positive-Unlabeled* ou PU), et faiblement supervisées. Les performances des modèles sous différents paradigmes sont visualisées à l'aide de diagrammes en boîte (Figures 5.1 et 5.2), qui illustrent la distribution des mesures de performance (AUCROC et AUCPR) pour les algorithmes de détection d'anomalies selon les différents paradigmes.

2. Voir Annex A pour les résultats détaillés des expériences menées dans la Partie II

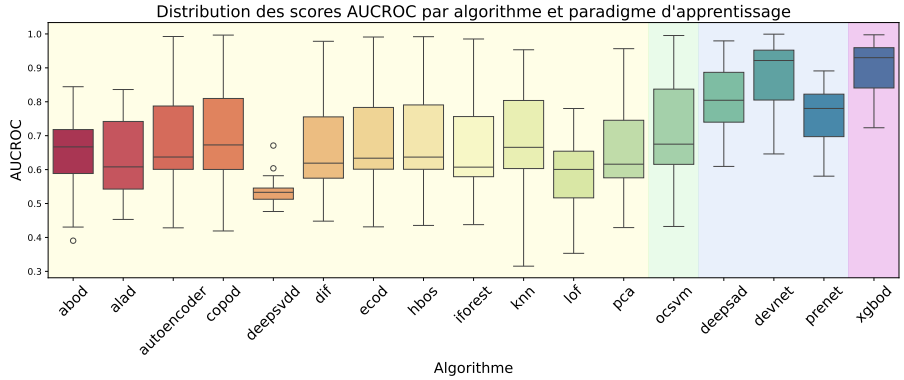


FIGURE 5.1 – Distribution des scores AUCROC des algorithmes de détection d'anomalies regroupés par paradigme d'apprentissage. Ce diagramme en boîte présente la distribution des scores AUCROC pour divers algorithmes de détection d'anomalies, catégorisés par leurs paradigmes d'apprentissage : non supervisé, semi-supervisé NO, semi-supervisé PU et faiblement supervisé. Les couleurs de fond dans la figure représentent ces paradigmes : jaune clair pour non supervisé, vert clair pour semi-supervisé NO, bleu clair pour semi-supervisé PU et violet clair pour faiblement supervisé.

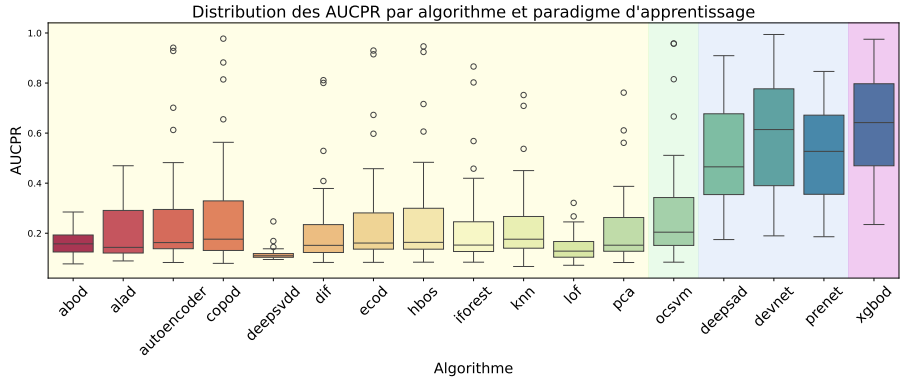


FIGURE 5.2 – Distribution des scores AUCPR des algorithmes de détection d'anomalies regroupés par paradigme d'apprentissage. Ce diagramme en boîte présente les scores AUCPR pour divers algorithmes de détection d'anomalies.

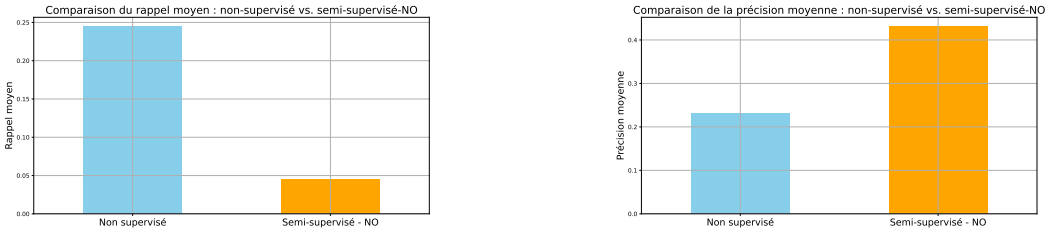
5.3.1.1 Paradigme non supervisé et paradigmes partiellement supervisés

Apprentissage non supervisé Historiquement, la détection d'anomalies a été abordée comme une tâche typiquement non supervisée, de même nature que le clustering. Aussi, l'apprentissage non supervisé constitue une grande partie des méthodologies traditionnelles de détection d'anomalies. Ces algorithmes s'appuient uniquement sur la structure intrinsèque des données et opèrent sans aucune entrée étiquetée. L'absence de données étiquetées peut entraver considérablement les performances, car ces modèles ne disposent pas des signaux contextuels fournis par les étiquettes qui définissent ce qui constitue une anomalie. En conséquence, ils ont souvent du mal à différencier efficacement les variations normales des véritables anomalies. Cette déficience est particulièrement prononcée dans les scénarios complexes qui exigent une compréhension nuancée des données anormales.

Les résultats des expériences soulignent ces défis, comme en témoignent les scores AUCPR et AUCROC inférieurs observés pour les algorithmes non supervisés par rapport à leurs homologues partiellement supervisés. En particulier, leurs scores AUCPR sont généralement les plus bas parmi tous les paradigmes, ce qui reflète des taux élevés de faux positifs et de faux négatifs. La précision et le rappel souffrent de l'absence de données étiquetées pour guider le modèle, conduisant à une mauvaise performance globale de détection. Les diagrammes en boîte dans les Figures 5.1 et 5.2 illustrent clairement ces tendances, montrant que les algorithmes non supervisés sont à la traîne en termes d'efficacité.

Apprentissage semi-supervisé NO Le paradigme semi-supervisé NO est souvent considéré comme la forme la plus classique d'apprentissage semi-supervisé dans la détection d'anomalies. En entraînant les modèles exclusivement sur des instances normales annotées, cette approche se concentre principalement sur la modélisation de ce qui constitue un comportement normal des données.

Ce type de modèle est particulièrement utile pour réduire le nombre de faux positifs, c'est-à-dire d'instances normales classées à tort comme des anomalies. Cependant, l'absence de données d'anomalies étiquetées limite la capacité du modèle à reconnaître et à comprendre les patterns anormaux, ce qui se traduit souvent par un taux plus élevé de faux négatifs. Cela peut être observé dans les résultats présentés dans les Figures 5.3a et 5.3b, où les méthodes semi-supervisées NO montrent une précision moyenne plus élevée sur l'ensemble des corpus par rapport aux méthodes non supervisées, bien que ce soit au détriment du rappel.



(a) Rappel (b) Précision

FIGURE 5.3 – Comparaison du rappel et de la précision pour les méthodes non supervisées et semi-supervisé NO.

En termes de scores AUCROC et AUCPR, le paradigme semi-supervisé NO, malgré des performances légèrement supérieures à celles des méthodes non supervisées, reste nettement inférieur aux paradigmes semi-supervisé PU et faiblement supervisé. Une amélioration tellement modeste par rapport aux méthodes non supervisées indique que la seule intégration d'échantillons normaux annotés peut contribuer à améliorer la détection d'anomalies dans les textes, mais son efficacité est très limitée.

Apprentissage semi-supervisé PU L'incorporation d'une petite quantité d'échantillons positifs étiquetés, comme le démontre le paradigme semi-supervisé PU, améliore considérablement la capacité d'un modèle à détecter des anomalies. Cette approche utilise des anomalies étiquetées pour fournir des informations essentielles sur la nature des anomalies, ce qui permet d'améliorer considérablement les taux

de vrais positifs et de renforcer les capacités de détection globales du modèle. L'apprentissage PU maîtrise habilement le compromis entre les faux positifs et les faux négatifs, montrant un équilibre plus efficace que l'approche semi-supervisée NO.

Comme en témoignent les mesures de performance illustrées dans les Figures 5.1 et 5.2, l'introduction des anomalies étiquetées a un impact profond. Avec seulement 40% des anomalies annotées dans les données d'entraînement (qui ont un ratio global d'anomalies de 10%), les méthodes semi-supervisées PU atteignent un AUCROC de 0,818 et un AUCPR de 0,537. Ces chiffres dépassent de manière significative ceux des méthodes non supervisées, qui obtiennent 0,647 en AUCROC et 0,215 en AUCPR, ainsi que des méthodes semi-supervisées NO, qui enregistrent 0,704 en AUCROC et 0,293 en AUCPR. Des améliorations aussi remarquables, notamment par rapport aux méthodes traditionnelles semi-supervisées NO, soulignent la pertinence des échantillons d'anomalies annotés. Elles suggèrent que même une portion modeste d'échantillons positifs étiquetés peut améliorer de manière significative les performances. Les anomalies s'avèrent beaucoup plus informatives que les données normales annotées.

Apprentissage faiblement supervisé L'apprentissage faiblement supervisé intègre à la fois des étiquettes normales et anormales, bien que de manière partielle, ce qui permet un processus d'apprentissage plus approfondi. La présence des deux types de données permet aux algorithmes d'établir des normes de comportement plus précise et de mieux définir les anomalies, apprenant ainsi une frontière de décision plus claire et réduisant à la fois les faux positifs et les faux négatifs. En introduisant 40% de données annotées dans le processus d'apprentissage, le paradigme faiblement supervisé a obtenu les meilleures performances globales parmi tous les paradigmes, avec une AUCROC moyenne de 0,9 et une AUCPR de 0,64 sur l'ensemble des corpus.

L'analyse démontre clairement que l'intégration de données partiellement étiquetées améliore significativement les performances des algorithmes de détection d'anomalies textuelles. Les paradigmes faiblement supervisés et PU semi-supervisés surpassent largement les paradigmes non supervisés et semi-supervisés NO. La progression des performances des méthodes non supervisées aux méthodes faiblement supervisées met en évidence l'impact crucial de l'intégration de données étiquetées, en particulier les anomalies véritables. Cela nous conduit à examiner plus en profondeur l'utilisation optimale des échantillons étiquetés : combien d'échantillons annotés sont nécessaires pour obtenir des améliorations notables, et quels types d'anomalies annotées sont les plus utiles pour entraîner les modèles ?

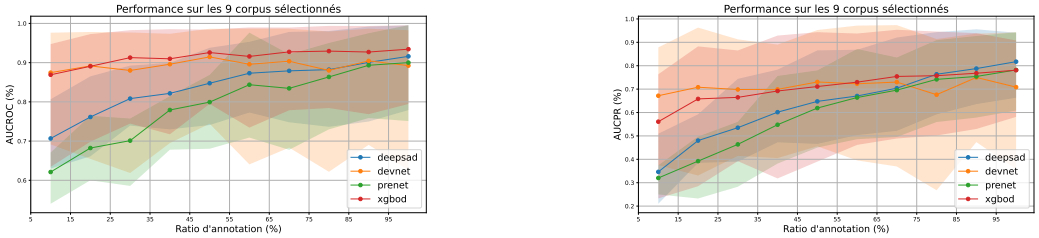
5.3.1.2 Ratio d'annotation optimal

Pour déterminer l'utilisation optimale des échantillons étiquetés dans la détection d'anomalies, nous avons mené des expériences pour analyser l'impact de diverses ratios d'annotation (γ_l) sur la performance des algorithmes. Les ratios d'annotations testés allaient de 10% à 100%, où 100% représente un scénario entièrement étiqueté similaire à une tâche de classification à classes ouvertes avec un déséquilibre extrême des classes.

Analyse des tendances générales

— **Augmentation des ratios d'annotation.** Une amélioration générale des performances est observée à mesure que le ratios d'annotation augmente de 10% à 100%. Cette tendance, clairement visible dans les deux figures de 5.4, confirme que l'augmentation de la quantité de données étiquetées améliore systématiquement les performances du modèle.

— **Plateau de performance.** Cependant, il y a un plateau de performance notable où des données étiquetées supplémentaires apportent des rendements décroissants. Ce phénomène se produit généralement au-delà du taux d'annotation de 50-60%. Comme illustré dans la Figure 5.5, en traçant le gain incrémental des mesures de performance par rapport au ratio d'annotation pour chaque algorithme et en définissant un seuil de rendement décroissant pour chaque mesure (par exemple, 0,05 pour AUCROC et 0,02 pour AUCPR), nous pouvons constater qu'au-delà de ce point (γ_l > 50%), aucun des algorithmes n'obtient d'améliorations significatives.



(a) AUCROC (b) AUCPR

FIGURE 5.4 – Changements de performance en AUCROC et AUCPR en fonction de ratios d'annotation. Ces figures illustrent les changements dans les métriques de performance (AUCROC et AUCPR) pour différents algorithmes de détection d'anomalies à mesure que le ratio d'annotation augmente. Elles tracent la progression de γ_l de 5% à 95%, démontrant comment ces métriques évoluent avec plus de données étiquetées. Les algorithmes présentés incluent DeepSAD, DevNet, PReNet et XGBOD, évalués sur plusieurs corpus.

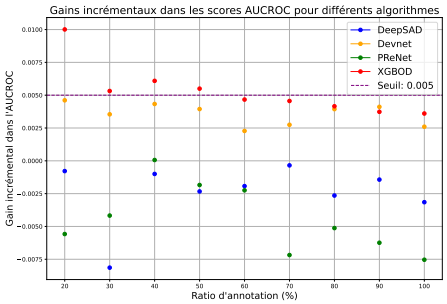
Observations spécifiques aux algorithmes

— Les algorithmes semi-supervisés PU comme PReNet et DevNET montrent des améliorations significatives des performances jusqu'à la marque d'annotation de 50%-60%, après laquelle leurs performances se stabilisent. Cela indique leur dépendance aux données étiquetées pour une détection efficace des anomalies.

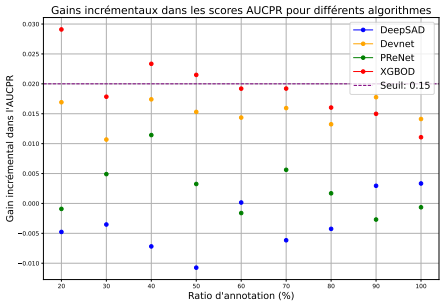
— XGBOD, une méthode faiblement supervisée, présente des gains de performance initiaux qui se stabilisent rapidement après 20% d'annotation pour l'AUCROC et 30% pour l'AUCPR.

— DevNet, en revanche, démontre des performances constamment élevées dès le départ, indiquant son efficacité avec de faibles quantités de données étiquetées et plus grande indépendance vis-à-vis de l'augmentation du volume des données annotées.

D'après notre analyse, le ratio d'annotation optimal semble se situer autour de 50%. Par conséquent, pour un ensemble de données d'entraînement contenant 1000 échantillons avec un taux d'anomalie estimé à 1%, l'annotation d'au moins 5 anomalies devrait permettre d'obtenir de bonnes performances. Ce repère fournit une



(a) Gain incrémental du score AUCROC



(b) Gain incrémental du score AUCRPR

FIGURE 5.5 – Gains incrémentaux des mesures de performance avec l'augmentation des annotations. Ces figures présentent les améliorations progressives des scores AUCROC et AUCPR lorsque γ_l passe de 20% à 100 %. Chaque figure met en évidence les seuils de performance spécifiques pour divers algorithmes de détection d'anomalies. Les lignes de seuil, fixées à des gains de 0,005 pour l'AUCROC et de 0,015 pour l'AUCPR, indiquent visuellement les points au-delà desquels des augmentations supplémentaires des données annotées ont un impact décroissant.

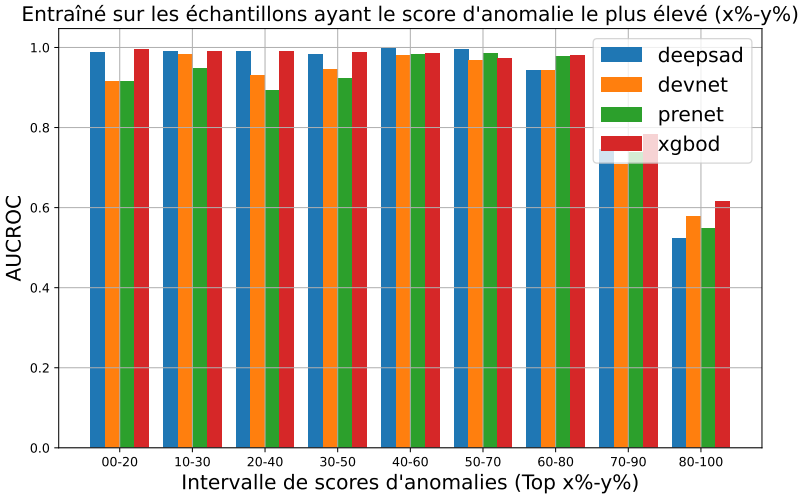
ligne directrice pratique pour déployer efficacement les ressources dans les tâches de détection d'anomalies textuelles, garantissant que les données annotées pour l'entraînement aient un impact maximal sur l'efficacité du modèle.

5.3.1.3 Sélection optimale des échantillons d'anomalies

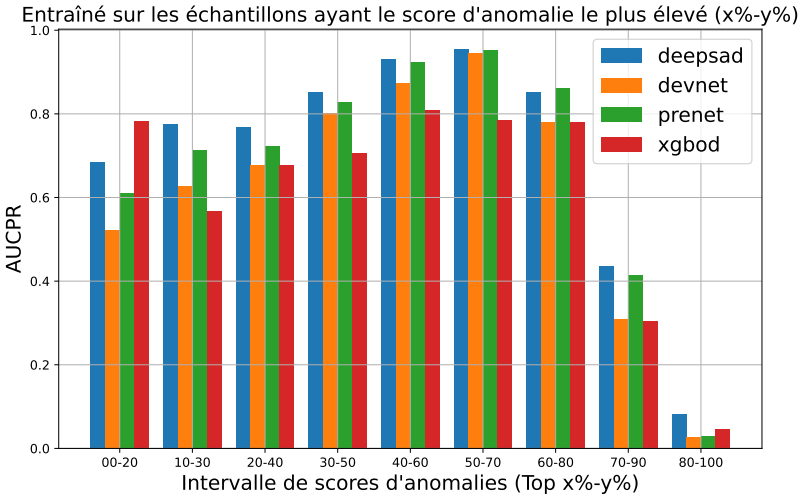
Afin d'explorer la sélection optimale des échantillons d'anomalies pour l'entraînement des modèles de détection d'anomalies, nous avons mené des expériences analysant la relation entre le degré d'anormalité des échantillons annotés et la performance des algorithmes. L'objectif est de déterminer si des échantillons fortement anormaux ou des échantillons plus nuancés et moins discernables sont plus efficaces pour définir la frontière de décision pour la détection d'anomalies.

Cette expérience s'est déroulée comme suit :

1. **Adaptation de corpus.** Nous avons adapté les corpus originaux (AGNews et 20NG) en classant tous les échantillons d'anomalies dans l'ensemble d'entraînement selon leurs scores d'anomalie. Ces scores ont été attribués en utilisant cinq modèles de détection d'anomalies non supervisés traditionnels : IForest, HBOS, KNN, LOF et DBSCAN. Les échantillons avec les scores les plus élevés étaient les plus anormaux, et ceux avec des scores plus bas étaient considérés comme moins discernables.
2. **Approche par fenêtre coulissante.** Une approche par fenêtre coulissante a été employée, où le ratio d'annotation $\gamma_l = 20\%$ a été utilisé comme taille de fenêtre. À chaque essai, les 20% d'échantillons les plus anormaux ont été sélectionnés comme données étiquetées, tandis que les 80% restants étaient non étiquetés. Le processus a commencé à partir des 20% d'échantillons les plus anormaux et s'est déplacé vers le bas par paliers de 10%, couvrant des intervalles de 0-20% à 80-100%.



(a) AUCROC



(b) AUCPR

FIGURE 5.6 – Impact de la sélection des échantillons d'anomalies sur la performance des modèles à travers différents intervalles de scores d'anomalie. Les figures affichent les scores AUCROC et AUCPR pour les modèles entraînés sur des échantillons d'anomalies provenant des corpus AGNews et 20NG, segmentés en divers intervalles de scores d'anomalie allant des plus anormaux (0-20%) aux moins anormaux (80-100%). Les graphiques montrent clairement comment la sélection des échantillons d'anomalies, basée sur leur niveau d'anomalité, affecte les performances de différents algorithmes de détection d'anomalies tels que DeepSAD, DevNet, PReNet et XGBOD.

Les modèles ont ensuite été entraînés à partir des échantillons d'anomalies étiquetés sélectionnés pour chaque intervalle et évalués en fonction de leurs scores AUCROC et AUCPR, comme le montre la Figure 5.6.

Analyse des tendances générales

— **Échantillons fortement anormaux.** Les modèles entraînés avec les échantillons les plus anormaux montrent généralement de meilleures performances initiales en termes de scores AUCROC, suggérant que ces échantillons sont essentiels pour aider les modèles à apprendre des frontières de décision distinctes.

— **Déclin des performances.** À mesure que l'entraînement passe à des intervalles avec des échantillons moins anormaux (par exemple, 80-100%), il y a un déclin notable des mesures de performance. Ces échantillons sont moins discernables des instances normales, ce qui rend plus difficile pour les modèles d'apprendre et de définir des frontières robustes.

Observations spécifiques aux algorithmes

— **Méthodes semi-supervisées PU (DeepSAD, DevNet, PReNet).** En termes de score AUCPR, lorsqu'ils sont entraînés avec les échantillons les plus anormaux (top 20%), ces modèles montrent une haute précision, car ces échantillons sont plus faciles à distinguer. Cependant, leur rappel est faible, car ils tendent à surajuster aux cas les plus extrêmes et échouent à généraliser aux instances moins anormales. Dans les intervalles de milieu de gamme, où l'ensemble d'entraînement inclut un mélange d'échantillons extrêmement et modérément anormaux, ces modèles atteignent un meilleur équilibre entre précision et rappel, conduisant à des scores AUCPR améliorés.

— **XGBOD.** En tant que méthode ensembliste, XGBOD gère plus efficacement une large gamme de scores d'anomalie, même lorsqu'il est entraîné avec les échantillons les plus anormaux, ce qui se traduit par des scores AUCPR constamment élevés dès le départ.

5.3.1.4 Erreurs d'annotation

Les méthodes faiblement supervisées et semi-supervisées sont particulièrement sensibles à la qualité des annotations. Pour étudier cette sensibilité, nous avons introduit différents niveaux d'erreurs d'annotation dans les ensembles d'entraînement en inversant les étiquettes de $\epsilon_a\%$ des échantillons, avec ϵ_a variant de 1 à 50. Notre analyse, menée sur DeepSAD, DevNet, PReNet et XGBOD, démontre comment ces erreurs d'annotation impactent les performances des modèles, comme le montrent les courbes ROC et les scores AUCROC fournis dans la Figure 5.7 et le Tableau 5.2.

Analyse des tendances générales Pour tous les algorithmes, on observe une nette tendance vers la baisse des scores AUCROC à mesure que le taux d'erreur d'annotation ϵ_a augmente. Cette tendance est un indicateur fort de la sensibilité générale à la qualité des annotations, avec les taux d'erreur plus élevés conduisant systématiquement à des performances réduites. La dégradation des performances des modèles devient plus prononcée à des taux d'erreur plus élevés, en particulier au-delà de 25%, suggérant un seuil au-delà duquel l'impact des erreurs compromet significativement l'efficacité des algorithmes de détection. Lorsque le taux d'erreur ϵ_a

Algorithme	Sans Err.	Err. 1%	Err. 5%	Err. 10%	Err. 25%	Err. 50%
DeepSAD	0,80	0,78	0,75	0,71	0,63	0,50
DevNet	0,90	0,89	0,88	0,86	0,66	0,49
PReNet	0,75	0,75	0,73	0,70	0,62	0,49
XGBOD	0,91	0,91	0,88	0,86	0,76	0,49

TABLE 5.2 – Scores AUCROC sous différents taux d'erreur d'annotation. Ce tableau présente les scores AUCROC obtenus par DeepSAD, DevNet, PReNet et XGBOD pour un spectre d'erreurs d'annotation allant de 0 à 50%, fournissant une mesure quantitative de l'impact de l'exactitude des annotations sur la performance des algorithmes.

atteint 50%, l'exactitude de la prédiction des modèles diminue à un niveau proche de la devinette aléatoire.

Observations spécifiques aux algorithmes

— **DeepSAD** (Figure 5.7a) manifeste une diminution progressive mais constante du score AUCROC. Le déclin systématique à travers tous les taux d'erreur suggère une sensibilité linéaire aux erreurs d'annotation, indiquant que la performance de DeepSAD diminue de manière prévisible à mesure que la quantité d'erreurs d'annotation augmente.

— **DevNet** (Figure 5.7b) commence avec un score élevé de 0,90 et maintient des performances relativement stables jusqu'à $\epsilon_a = 10$. Cependant, au-delà de ce point, il y a une chute brutale, en particulier de 0,86 à $\epsilon_a = 10$ à 0,66 à $\epsilon_a = 25$, ce qui constitue le déclin de performance le plus prononcé. Cela indique que, bien que DevNet soit robuste contre les inexactitudes mineures, il rencontre des difficultés significatives avec des taux d'erreur plus élevés.

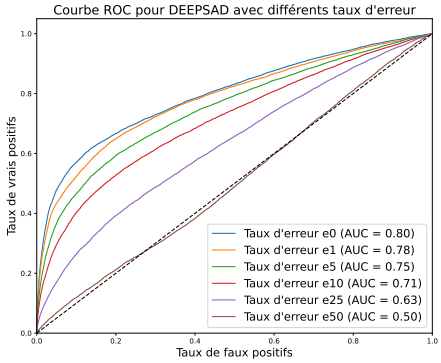
— **PReNet** (Figure 5.7c) suit la tendance générale, en maintenant des performances relativement stables jusqu'à $\epsilon_a = 10$. Au-delà de ce point, il y a une forte baisse de performance, s'alignant avec les tendances observées dans les autres modèles.

— **XGBOD** (Figure 5.7d) démontre la plus grande résilience aux faibles taux d'erreur, en maintenant des performances élevées jusqu'à $\epsilon_a = 10$, similaire à DevNet. Cependant, contrairement à DevNet, XGBOD subit une dégradation plus légère des performances au-delà de ce point, indiquant une plus grande robustesse face à l'augmentation des erreurs d'annotation.

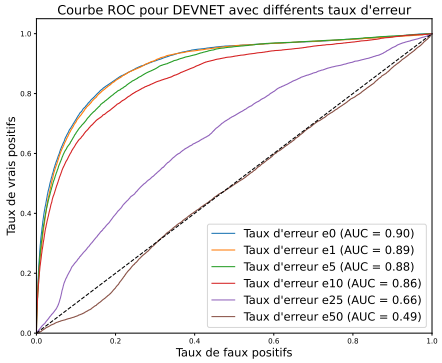
5.3.2 Nature d'anomalie textuelle

Dans notre étude, nous avons également exploré la compatibilité et l'efficacité des algorithmes de fouille de données pour trois types d'anomalies textuelles pertinentes dans le domaine de veille : les anomalies thématiques, les sentiments déviants et les discours de haine (voir le Tableau 5.3 et la Figure 5.8). Ces anomalies ont été analysées à l'aide de corpus adaptés à partir de jeux de données initialement conçus pour la classification de textes/thématiques, l'analyse des sentiments et la détection des discours de haine.

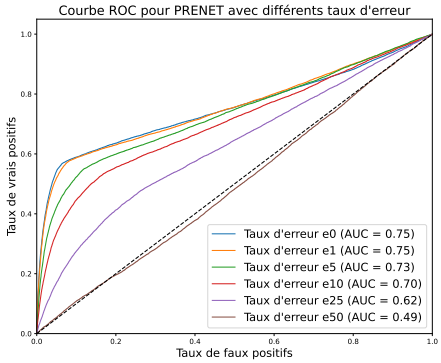
De manière générale, la performance des algorithmes de détection d'anomalies peut être influencée par l'origine des jeux de données, qui varient souvent considérablement en termes de qualité de texte et de registre de langue. Par exemple, les



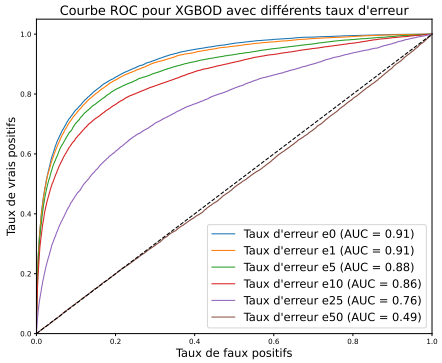
(a) DeepSAD



(b) DevNet



(c) PReNet



(d) XGBOD

FIGURE 5.7 – Courbes ROC pour les algorithmes de détection d'anomalies avec des taux d'erreur variables. Cette série de courbes ROC illustre les performances de DeepSAD, DevNet, PReNet et XGBOD sous différents taux d'erreur d'annotation, soulignant comment les erreurs influencent les capacités de détection de chaque modèle.

Type	**Min**	**Max**	**Moyenne**	**Médiane**	**Écart type**
Anomalies thématiques	*0,3153*	0,9975	0,7186	0,6620	0,1635
Sentiments déviants	0,4403	0,9822	0,6955	0,6807	0,1279
Discours de haine	0,3904	0,9399	0,5268	0,4765	0,1386

TABLE 5.3 – Statistiques de performance des modèles de détection d'anomalies pour différents types d'anomalies. Ce tableau présente les statistiques de performance (minimum, maximum, moyenne, médiane et écart type des scores AUCROC) des algorithmes de détection d'anomalies appliqués à différents types d'anomalies textuelles.

anomalies thématiques proviennent généralement de sources telles que la presse ou des contenus élaborés de manière professionnelle, où la langue est standard et le texte bien structuré. En revanche, les anomalies de sentiment proviennent souvent de grands sites web et applications hébergeant des avis ou commentaires d'utilisateurs. Dans ces contextes, la langue, bien que moins raffinée que celle de la presse, reste plus structurée que celle utilisée dans les réseaux sociaux. Les discours de

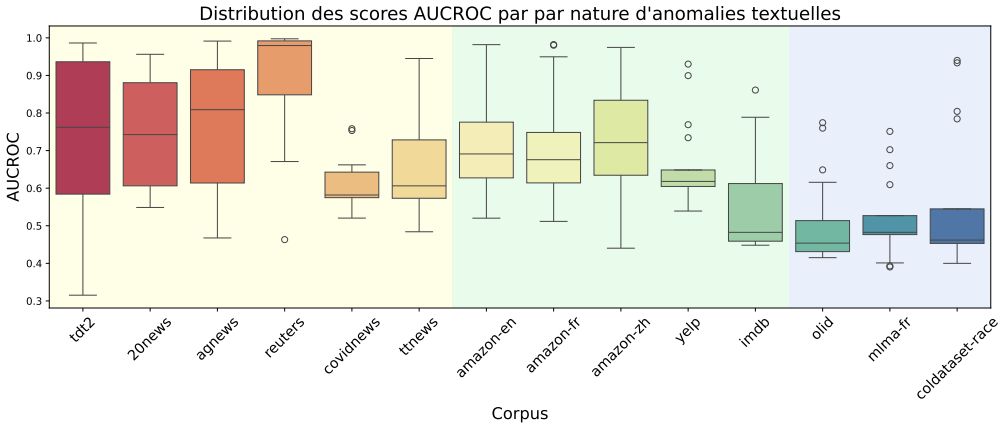


FIGURE 5.8 – Distribution des performances des algorithmes pour différents types d'anomalies textuelles. Cette figure illustre la distribution des scores AUCROC pour les algorithmes de détection d'anomalies à travers trois types d'anomalies textuelles : les anomalies thématiques, les sentiments déviants et les discours de haine, représentés respectivement par le jaune clair, le vert clair et le bleu clair.

haine, quant à eux, se trouvent principalement sur les plateformes de réseaux sociaux ou les forums, où la langue est souvent très familière et contient des expressions argotiques et non standard. La nature informelle et variée de la langue sur ces plateformes rend l'identification du langage offensant particulièrement difficile avec une simple combinaison d'algorithmes de fouille de données et de techniques de représentation de texte.

Anomalies thématiques Les algorithmes se révèlent plus efficaces pour identifier les anomalies aux niveaux thématiques, à savoir les sujets nouveaux ou inconnus, comme en témoigne un score AUCROC moyen d'environ 0,72. Cette performance élevée est probablement due à la similarité entre la détection d'anomalies thématiques et les tâches traditionnelles basées sur les thématiques. En effet, des techniques comme la modélisation thématique et le clustering partagent des principes avec les méthodes utilisées pour la détection d'anomalies, telles que celles basées sur la proximité, les statistiques, et la reconstruction. Ainsi, dans un système de veille de l'actualité centré sur des sujets spécifiques, les algorithmes peuvent efficacement détecter et signaler les événements nouveaux ou inattendus comme des anomalies. Cette aptitude repose sur la capacité des modèles à identifier les déviations par rapport aux modèles thématiques établis, en utilisant des mécanismes similaires à ceux de la modélisation thématique et du clustering.

Toutefois, les performances ne sont pas uniformément élevées pour tous les jeux de données, comme l'indique le score minimum de l'AUCROC de 0,32. Ce score exceptionnellement bas est dû à l'inclusion de multiples corpus complémentaires où les distinctions entre exemples normaux et anormaux sont subtiles et complexes. Par exemple, différencier entre « crise économique (normal) » et « violence contre le gouvernement pendant la crise (anomalie) » ou entre conflit israélo-palestinien » et « négociations de paix entre Israël et la Palestine » pose des défis significatifs. Dans de tels contextes, les nuances subtiles entre ce qui est considéré comme normal et ce qui est signalé comme anormal peuvent gravement affecter l'efficacité des algorithmes de fouille de données. Cette variabilité met en évidence une limitation critique des

approches actuelles de détection d'anomalies, soulignant le besoin de méthodologies plus nuancées et sensibles au contexte, capables de discerner des variations légères mais significatives dans les données thématiques.

Sentiment déviant Le score moyen de l'AUCROC pour la détection des sentiments déviants est d'environ 0,70, avec une plage de valeurs allant de 0,44 à 0,98. Cela reflète une performance modérée, avec une variabilité moins extrême par rapport aux anomalies thématiques.

D'une part, identifier des anomalies liées aux sentiments, comme les changements de sentiment, présente un défi en raison des façons complexes et diversifiées dont les émotions sont exprimées dans les textes. Une détection efficace nécessite des algorithmes capables non seulement d'interpréter les expressions explicites, mais aussi de comprendre les implications nuancées telles que le sarcasme et d'autres indices contextuels subtils. D'autre part, bien que l'analyse inclue des cas moins distincts, la dégradation de la performance observée n'est pas aussi marquée que pour les anomalies thématiques. Cela suggère que, bien que la détection des anomalies de sentiment présente ses propres défis, ceux-ci sont généralement moins sévères que ceux rencontrés lors de la différenciation entre des thématiques étroitement liées mais distinctes.

Discours de haine L'application des algorithmes de détection d'anomalies pour identifier les discours de haine et le langage offensant présente des défis significatifs, comme en témoigne un score AUCROC moyen d'environ 0,53 et une médiane de 0,48, proche d'une estimation arbitraire. Ce défi résulte non seulement de la nécessité de saisir les expressions implicites nuancées, comme c'est le cas pour détecter les sentiments déviants, mais aussi de l'obligation de comprendre les contextes sociaux et culturels complexes intégrés dans la langue. L'approche actuelle utilise une combinaison des algorithmes de fouille de données avec des techniques de transformation de texte, ce qui rend la compréhension de la langue entièrement dépendante de la capacité de ces techniques de transformation à capturer les connaissances linguistiques, y compris les éléments lexicaux, syntaxiques et sémantiques. Cette dépendance peut s'avérer problématique lorsqu'il s'agit de traiter des anomalies plus complexes telles que le langage offensant, où la compréhension d'expressions nuancées et culturellement contextuelles est cruciale.

L'analyse des performances des algorithmes à travers différents types d'anomalies met en évidence la nécessité de trouver des approches adaptées aux différents défis posés par les données textuelles. Alors que les techniques actuelles peuvent gérer efficacement les anomalies thématiques, des anomalies plus complexes telles que les sentiments déviants et les discours de haine exigent l'intégration de la compréhension contextuelle et de techniques TALN avancées pour une détection exacte.

5.3.3 Techniques de représentation

Pour analyser l'impact des différentes techniques de représentation sur la performance de la détection d'anomalies dans le texte, nous effectuons deux comparaisons principales : 1) Représentations traditionnelles (TFIDF) versus modèles de langue pré-entraînés (Sentence-BERT ou SBERT) ; 2) Modèles SBERT monolingues versus modèles SBERT multilingues.

5.3.3.1 TFIDF contre Sentence-BERT

Analyse de performance Nos résultats (Tableau 5.4 et Figure 5.9) montrent que les modèles SBERT surpassent systématiquement et considérablement TFIDF dans tous les jeux de données, avec une moyenne de AUCROC de 0,691. Dans certains cas, SBERT atteint des performances quasi parfaites (AUCROC proche de 1) dans le quartile supérieur de sa distribution. En revanche, de telles performances élevées ne sont généralement observées que dans des cas aberrants pour TFIDF.

	TFIDF	SBERT
Min	0,1707	0,2306
Max	0,9956	0,9998
Moyenne	0,5528	0,6910
Médiane	0,5173	0,6718
Écart-type	0,1563	0,1722

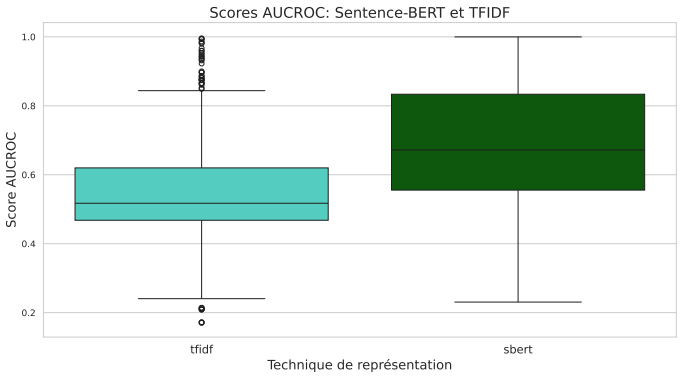


TABLE 5.4 – Statistiques comparatives de TFIDF et SBERT à travers les jeux de données et les algorithmes. Ce tableau fournit une analyse statistique détaillée des performances des techniques de représentation TFIDF (traditionnelle) et Sentence-BERT (avancée) sur différents jeux de données et algorithmes. Les performances sont quantifiées en termes de minimum, maximum, moyenne, médiane et écart-type des scores AUCROC.

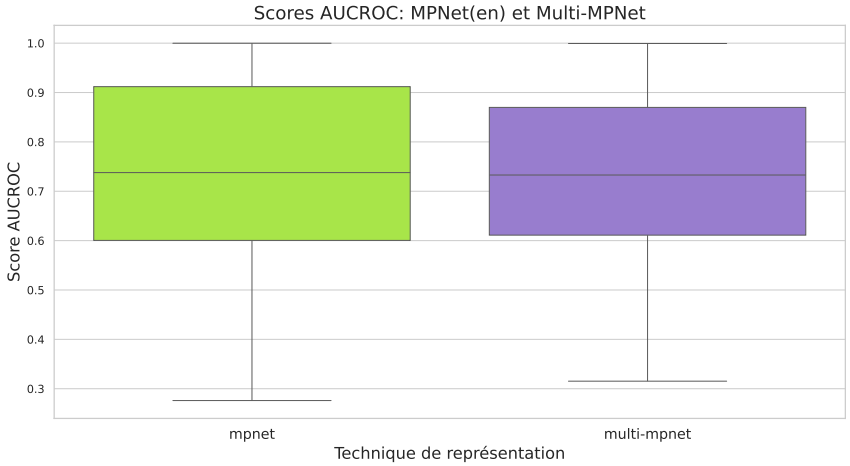
FIGURE 5.9 – Comparaison des performances de TFIDF et de SBERT à travers les jeux de données et les algorithmes. Cette figure illustre les scores AUCROC obtenus par les techniques de représentation traditionnelle TFIDF et avancée Sentence-BERT sur divers jeux de données et pour l'ensemble des algorithmes évalués

Compromis de performance Bien que SBERT offre des performances optimisées, c'est au prix de temps d'exécution accrus, ce qui est problématique dans les applications du monde réel. Le processus d'encodage de texte avec SBERT est relativement chronophage, dépassant parfois le temps nécessaire au processus d'inférence lui-même. En outre, les représentations vectorielles denses de SBERT peuvent ralentir la phase de détection, en particulier pour les modèles d'apprentissage profond, en raison de l'augmentation du temps de traitement.

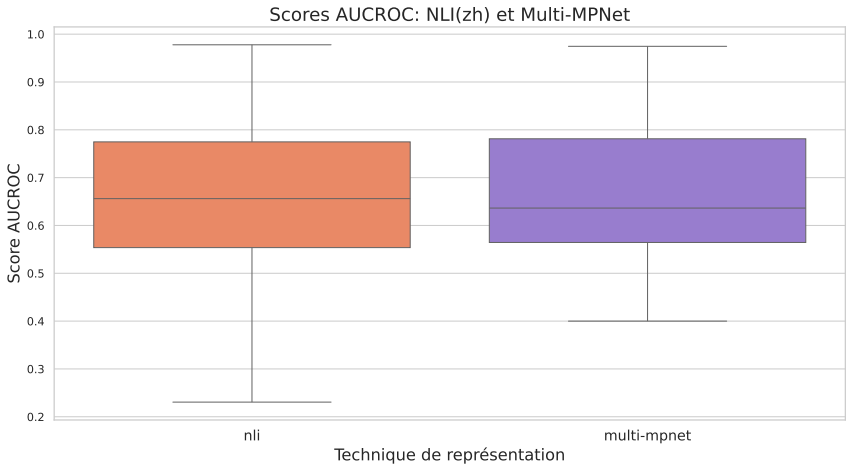
5.3.3.2 Modèles SBERT monolingues et modèles SBERT multilingues

Modèles en anglais et en chinois Les modèles monolingues pour le chinois (Figure 5.10a) et l'anglais (Figure 5.10b) montrent des performances comparables ou légèrement meilleures que les modèles multilingues, ce qui suggère des avantages dans la capture des nuances spécifiques à chaque langue.

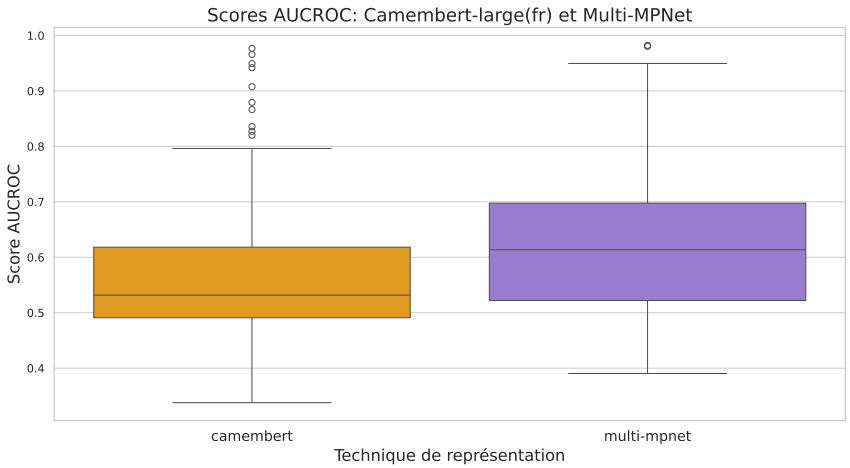
Modèle en français En revanche, le modèle français (Figure 5.10c) s'avère moins compétitif par rapport au modèle multilingue dans nos expériences. Cela pourrait



(a) Performance du modèle multilingue par rapport au modèle anglais sur les corpus anglais



(b) Performance du modèle multilingue par rapport au modèle chinois sur les corpus chinois



(c) Performance du modèle multilingue par rapport au modèle français sur les corpus français

FIGURE 5.10 – Comparaison des performances des modèles SBERT multilingues et monolingues sur des corpus spécifiques à chaque langue. Ces figures présentent une comparaison des scores AUCROC des modèles SBERT multilingues et monolingues sur des corpus anglais, chinois et français, respectivement, mettant en évidence leurs performances dans les tâches de détection d'anomalies.

être attribué à la dimensionnalité plus élevée (1024 dimensions) du modèle français, comparée aux 768 dimensions du modèle multilingue et des autres modèles monolingues. La dimensionnalité élevée est connue pour son impact négatif sur les performances en raison de la malédiction de la dimensionnalité, qui augmente la complexité et les exigences en matière de calcul.

En résumé, les représentations contextuelles denses comme SBERT offrent généralement de meilleures performances pour la détection d'anomalies par rapport aux méthodes traditionnelles comme TFIDF. Cependant, il est essentiel, pour les applications pratiques, de tenir compte de la dimensionnalité et des exigences computationnelles. Il est envisageable de parvenir aux résultats optimaux en employant des modèles aux dimensions appropriées qui équilibrent la capture sémantique riche et l'efficacité computationnelle.

5.3.4 Calcul des scores

Le mécanisme de calcul des scores d'anomalie, autrement dit la base théorique sous-jacente, est le noyau d'un algorithme de détection d'anomalies. Par conséquent, notre analyse vise également à vérifier si certains types de scores d'anomalie offrent des avantages par rapport à d'autres et, le cas échéant, à élucider les raisons de leur efficacité. Comme illustré dans la Figure 5.11, les algorithmes sont classés en fonction de leur base théorique ou de leur mécanisme de calcul des scores d'anomalie, et leurs performances sont visualisées à l'aide d'un diagramme en boîte des scores AUCROC à travers divers corpus.

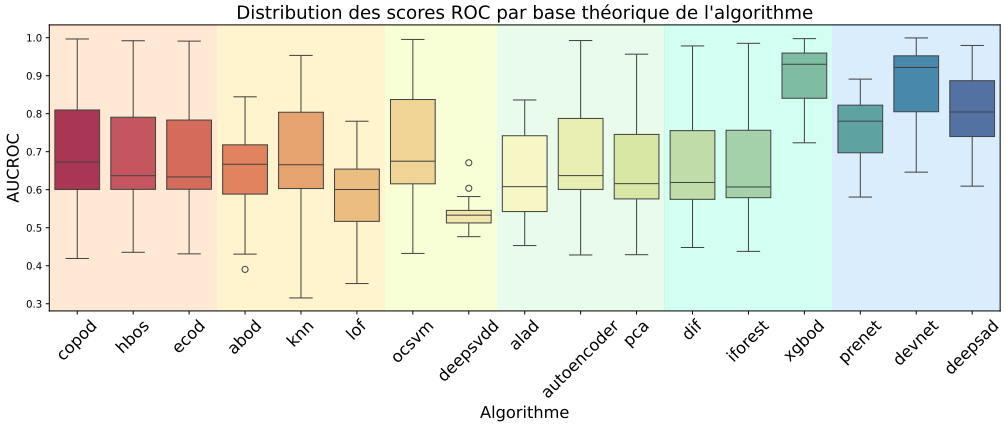


FIGURE 5.11 – Comparaison des performances des algorithmes de détection d'anomalie en fonction de la base théorique. Cette figure montre les scores AUCROC obtenus par divers algorithmes à travers différents corpus. Les algorithmes sont classés en fonction de la base théorique ou du mécanisme utilisé pour le calcul des scores d'anomalie, avec les couleurs de fond indiquant : rouge clair pour les algorithmes basés sur les statistiques/probabilités, orange clair pour les algorithmes basés sur la proximité, jaune clair pour les algorithmes basés sur le domaine, vert clair pour les algorithmes basés sur la reconstruction, bleu clair pour les algorithmes basés sur l'apprentissage ensembliste, et bleu pour les mécanismes récemment proposés (basés sur l'apprentissage des caractéristiques d'anomalie et basés sur l'apprentissage du score d'anomalie).

Algorithmes à base de proximité et à base de statistiques Les algorithmes de ce groupe, qu'ils soient traditionnels (comme KNN, LOF, HBOS) ou récemment proposés (comme COPOD, ECOD), ne parviennent pas à offrir un avantage de performance statistiquement significatif par rapport aux autres types d'algorithmes. L'absence de performances nettement supérieures indique que ces méthodes ne permettent pas nécessairement de saisir efficacement les patterns complexes inhérents aux données textuelles. Leur recours exclusif à la proximité ou aux propriétés statistiques peut s'avérer inadéquat pour les tâches de détection d'anomalies textuelles nuancées. De plus, les algorithmes basés sur la proximité et les statistiques souffrent de la « malédiction de la dimensionnalité », où leur performance se détériore à mesure que la dimensionnalité du jeu de données augmente. Bien que les méthodes les plus récentes, telles que COPOD et ABOD, soient conçues pour résoudre les problèmes liés à la haute dimension des données tabulaires, leur efficacité reste limitée lorsqu'elles sont appliquées aux données textuelles, qui sont denses et de vaste dimension. Les améliorations apportées par ces techniques récentes sont insuffisantes pour surmonter les défis posés par la grande dimensionnalité des données textuelles, ce qui conduit souvent à des résultats sous-optimaux de détection d'anomalies.

Algorithmes à base de reconstruction Les algorithmes à base de reconstruction, tels que les autoencodeurs et la PCA, ont connu un regain de popularité avec l'avènement des GANs au cours de la dernière décennie. Ces méthodes ont remporté un succès considérable dans diverses tâches de détection d'anomalies, notamment dans des domaines comme le traitement des images. Cependant, dans le contexte de la détection d'anomalies textuelles, elles tendent à produire des résultats médiocres pour plusieurs raisons. Premièrement, les paires nuancées introduisent du bruit : ces méthodes sont extrêmement sensibles au bruit, qui est courant dans les données textuelles. Des changements subtils de thématiques ou de sentiments peuvent créer du bruit que ces méthodes ont du mal à gérer. Cette sensibilité conduit à des performances exceptionnellement faibles pour les algorithmes comme les autoencodeurs et ALAD sur notre corpus complémentaire composé de paires normales-anomalies peu distinguables. Deuxièmement, les données textuelles sont intrinsèquement plus complexes que les images ou les caractéristiques numériques. Les méthodes à base de reconstruction réduisent généralement les données à une représentation de dimension inférieure et les reconstruisent ensuite pour identifier les déviations, mais les données textuelles ont des significations sémantiques complexes, et il s'avère difficile de saisir exactement ces nuances au cours de la reconstruction. Enfin, les dépendances contextuelles dans les textes ajoutent une couche de complexité supplémentaire. Un mot rare peut être normal dans un contexte mais une anomalie dans un autre. Capturer ces dépendances contextuelles est un défi pour les méthodes basées sur la reconstruction.

Algorithmes à base d'apprentissage ensembliste Des modèles comme IForest et DeepIForest, qui intègrent plusieurs détecteurs de bases homogènes, tendent à donner des résultats médiocres lors de la détection d'anomalies dans les textes. Cette dépendance envers une approche unique limite non seulement leur capacité à capturer un grand éventail de caractéristiques d'anomalies, mais peine également à cerner avec exactitude les variations à l'intérieur de la norme. En conséquence, cela entraîne deux problèmes majeurs : le rejet de certaines anomalies spécifiques qui s'écartent des caractéristiques d'anomalies identifiables et la production de faux po-

sitifs en raison de l'incapacité à discerner les nuances des variations normales. Ces défauts montrent les défis posés par l'utilisation de mécanismes de détection homogènes pour gérer les complexités des données textuelles, où les anomalies et les variations normales sont complexes et délicates.

En revanche, des algorithmes comme XGBOD, qui utilisent un mélange de différents détecteurs de base non supervisés, ont démontré des résultats impressionnants. Ce succès est attribué à leur capacité à exploiter diverses mécanismes de détection, chacune apportant des perspectives uniques sur les données. En incorporant divers estimateurs non supervisés, XGBOD capture un large spectre de caractéristiques d'anomalies, ainsi que des variations subtiles de la norme, améliorant ainsi à la fois sa robustesse et l'exhaustivité de la détection. En outre, l'intégration de différentes approches lui permet de se généraliser plus efficacement à travers divers jeux de données, compensant ainsi les faiblesses d'un détecteur unique lorsqu'il ne parvient pas à identifier un type d'anomalie spécifique.

Mécanisme récentes Les théories de base récemment proposées pour la détection, telles que l'apprentissage des scores d'anomalie de bout en bout (implémentés dans des modèles comme DevNet et PReNET) et les méthodes basées sur l'apprentissage des caractéristiques d'anomalie (utilisées dans DeepSAD), ont fait preuve d'une efficacité exceptionnelle dans les scénarios où les échantillons sont partiellement annotés. Ces techniques de scoring innovantes sont particulièrement aptes à maximiser l'utilisation de données annotées limitées. En apprenant et en affinant de manière adaptative les scores d'anomalie à partir des annotations disponibles, ces modèles peuvent profiter même d'un petit nombre d'anomalies annotées pour améliorer considérablement la qualité de la détection. De plus, ces mécanismes de scoring permettent une interaction dynamique avec les données, ce qui permet aux modèles à s'adapter continuellement aux changements dans le jeu de données au fil du temps.

Approches hybrides Certains algorithmes utilisent plusieurs théories fondamentales pour calculer les scores d'anomalie, maximisant ainsi efficacement les forces de diverses approches. XGBOD, par exemple, réunit l'apprentissage ensembliste avec l'apprentissage de représentation d'anomalie (augmentation de caractéristiques), ce qui lui permet de bénéficier simultanément de capacités de détection robustes et d'une représentation enrichie des données. De même, PReNet associe le calcul de la proximité à un mécanisme basé sur l'apprentissage des scores de bout en bout, offrant une approche globale de la détection d'anomalies. Ces approches hybrides enrichissent le cadre analytique en exploitant les forces de plusieurs théories de base, aidant ainsi à surmonter les limitations inhérentes à une approche singulière.

5.3.5 Efficacité du temps et des ressources

5.3.5.1 Apprentissage automatique et apprentissage profond

Dans l'application industrielle de la détection d'anomalies, il est crucial de trouver un équilibre optimal entre la performance de détection et l'efficacité en termes de temps et de ressources. À cet égard, déployer ou non des modèles d'apprentissage profond (*Deep Learning*, DL) devient une décision stratégique critique, car les méthodes DL nécessitent généralement plus de ressources informatiques et plus de temps par rapport aux modèles traditionnels d'apprentissage automatique (*Machine*

Learning, ML). Pour évaluer ces compromis, notre analyse compare la performance et l'efficacité entre les modèles ML et DL. Afin de garantir une comparaison relativement équitable et de minimiser les facteurs confondants, nous comparons des paires d'algorithmes qui reposent sur des bases théoriques similaires et qui ont la même disponibilité de données étiquetées. Par exemple, pour les modèles à base d'apprentissage ensembliste, IForest est comparé à DeepIForest. Étant donné que les modèles basés sur la proximité et les statistiques sont essentiellement peu profonds, notre attention se portera sur des modèles plus complexes tels que les méthodes basées sur la reconstruction, le domaine et l'apprentissage ensembliste. Les résultats de ces comparaisons sont détaillés dans le Tableau 5.5. De plus, une comparaison plus générale englobant tous les algorithmes est également effectuée, comme illustrés dans la figure 5.12.

Moyenne	**DL**	**ML**
Recons.	0.6560	0.6502
Domaine	*0.5354*	0.7037
Ensemble	0.6566	0.6567
Médiane	**DL**	**ML**
Recon.	0.6253	0.6160
Domaine	*0.5330*	0.6750
Ensemble	0.6190	0.6073

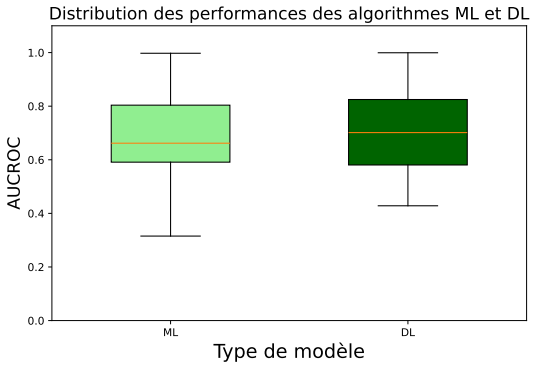


TABLE 5.5 – Comparaison des performances des modèles ML et DL dans la détection d'anomalies. Ce tableau fournit une comparaison statistique des modèles traditionnels d'apprentissage automatique (ML) et d'apprentissage profond (DL) à travers trois types d'algorithmes de détection d'anomalies : basé sur la reconstruction, basé sur le domaine et basé sur l'apprentissage ensembliste. Il présente les scores AUCROC moyens et médians obtenus par ces modèles sur plusieurs ensembles de données.

FIGURE 5.12 – Comparaison des performances des modèles ML et DL dans la détection d'anomalies. Cette figure illustre la distribution des scores AUROC pour les modèles traditionnels d'apprentissage automatique (ML) et d'apprentissage profond (DL) dans divers scénarios de détection d'anomalies. Pour les modèles ML, le score AUROC moyen est de 0,68577 et la médiane de 0,66185. Quant aux modèles DL, ils démontrent un score AUROC moyen légèrement plus élevé de 0,70847 et une médiane de 0,7014.

Les résultats de nos expériences révèlent que les apports des modèles DL dans la détection d'anomalies dans les textes sont modérés et incohérents. Cette observation est particulièrement vraie dans les cadres théoriques traditionnels et les paradigmes non supervisés. Par exemple, pour les méthodes à base de reconstruction, les modèles DL montrent un avantage minime sur les modèles ML, avec des scores moyens AUCROC de 0,6560 pour DL comparés à 0,6502 pour ML. Par contre, les modèles DL sont nettement sous-performants pour les méthodes à base de domaine, avec un AUROC moyen de seulement 0,5354, bien inférieur aux 0,7037 atteints par les modèles ML.

Dans de tels scénarios, les réseaux de neurones n'offrent pas systématiquement des améliorations pour la détection d'anomalies textuelles. Étant donné les exigences computationnelles plus élevées de DL, les modèles ML représentent souvent une option plus viable pour les tâches de détection d'anomalies textuelles non supervisées. Néanmoins, dans les cadres théoriques plus récents, notamment ceux qui concernent des scénarios faiblement ou semi-supervisés, les réseaux neuronaux s'avèrent essentiels pour utiliser efficacement des échantillons partiellement annotés.

Les analyses effectuées soulignent l'importance de choisir le bon modèle adapté au scénario d'application spécifique et à la disponibilité des données, garantissant ainsi une performance et une efficacité optimales dans les tâches de détection d'anomalies textuelles. À un stade ultérieur, l'analyse se penchera davantage sur des algorithmes spécifiques plutôt que sur des architectures générales de modèles, dans le but d'identifier le meilleur compromis entre l'efficacité temporelle et la performance de détection.

5.3.5.2 Efficacité de temps et performance de détection

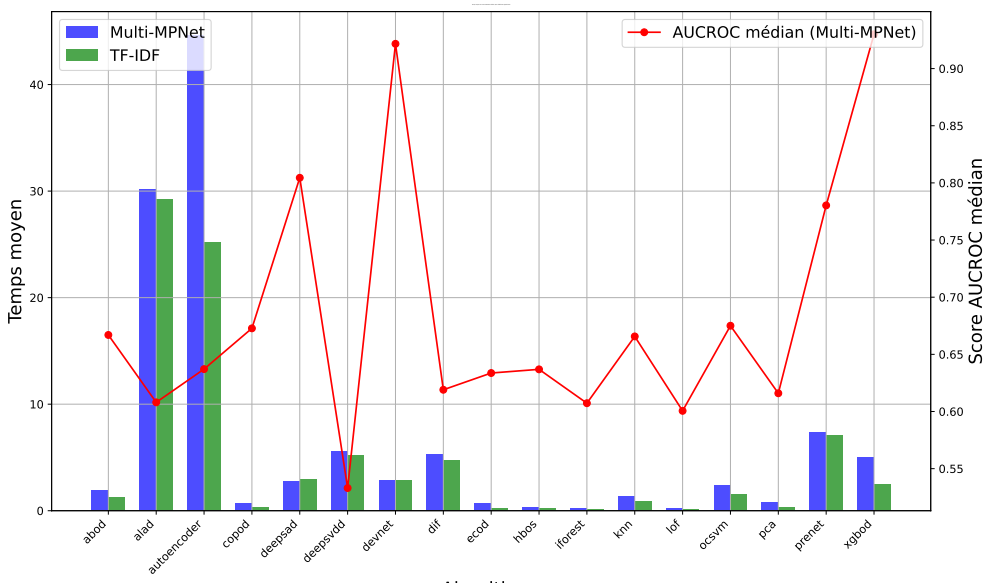


FIGURE 5.13 – Efficacité de temps par rapport à la performance de détection pour les algorithmes de détection d'anomalies. Cette figure présente une comparaison à double axe des algorithmes de détection d'anomalies, contrastant le temps moyen requis pour traiter mille documents (représenté par des barres) avec les médians des scores AUCROC (représentés par un graphique linéaire).

Aperçu général Pour donner un aperçu général de l'efficacité de temps et de la performance de détection pour chaque algorithme, nous avons tracé la performance AUCROC par rapport au temps requis pour traiter mille documents, comme illustré dans la Figure 5.13.

Méthodologie d'analyse Pour mieux comprendre le compromis entre performance et efficacité temporelle, nous avons procédé aux étapes analytiques suivantes :

1. **Normalisation des résultats de l'évaluation** : Normaliser les résultats relatives au temps et au score AUCROC afin d'assurer la comparabilité. La normalisation est effectuée en soustrayant la valeur minimale et en la divisant par la plage de l'ensemble de données :

$$\text{Temps normalisé} = \frac{\text{Temps} - \text{Temps}_{\text{min}}}{\text{Temps}_{\text{max}} - \text{Temps}_{\text{min}}}$$

$$\text{AUCROC Normalisé} = \frac{\text{AUCROC} - \text{AUCROC}_{\text{min}}}{\text{AUCROC}_{\text{max}} - \text{AUCROC}_{\text{min}}}$$

2. **Visualisation de l'équilibre temps-performance** : Créez un diagramme de dispersion où les abscisses représentent le temps et les ordonnées représentent le score AUCROC (Figure 5.14a). Ainsi, chaque point représente la performance d'un algorithme spécifique sur un corpus spécifique.
3. **Identification du front de Pareto** : Mettre en évidence les algorithmes sur le front de Pareto (Figure 5.14b). Les points situés sur le front de Pareto correspondent aux algorithmes qui atteignent le meilleur équilibre, ce qui signifie qu'aucun autre algorithme ne peut fournir une meilleure performance sans sacrifier plus de temps ou vice versa. Essentiellement, ces points indiquent les solutions les plus efficaces, où toute tentative d'amélioration d'un aspect conduirait à un compromis disproportionné dans l'autre.

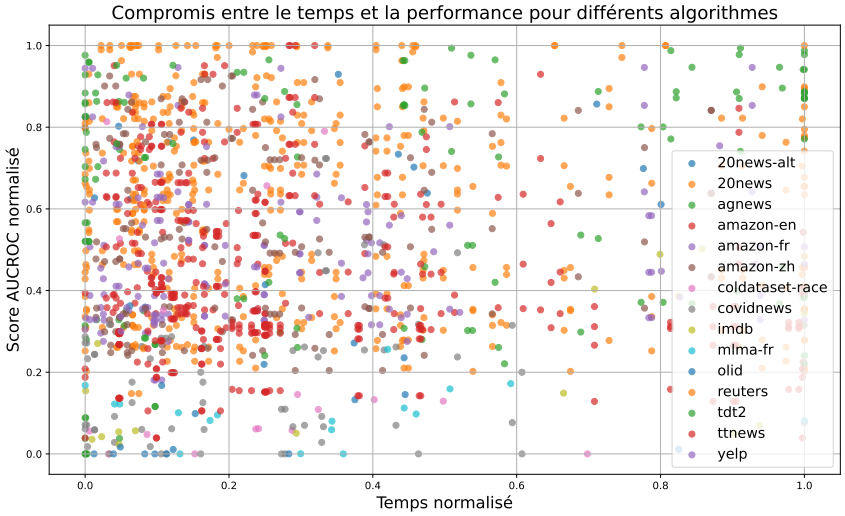
   Pour déterminer le front de Pareto, il convient d'analyser l'ensemble des choix possibles. Un point est sur le front de Pareto si aucun autre point ne le domine. Un point A domine B si A est au moins aussi bon que B dans toutes les dimensions et meilleur dans au moins une dimension. Mathématiquement, pour chaque point (x_A, y_A) sur le graphique, il se trouve sur le front de Pareto s'il n'y a pas de point (x_B, y_B) tel que $xB \leq xA$ et $yB \geq yA$ avec au moins une inégalité stricte.
4. **Calcul du score composite** : Calculer un score composite pour chaque algorithme afin de les classer en fonction de leur efficacité et de leur performance. Le score composite peut être défini en utilisant la formule :

$$\text{Score composite} = p_{\text{t}} \times (1 - \text{temps normalisé}) + p_{\text{p}} \times \text{AUCROC normalisé}$$

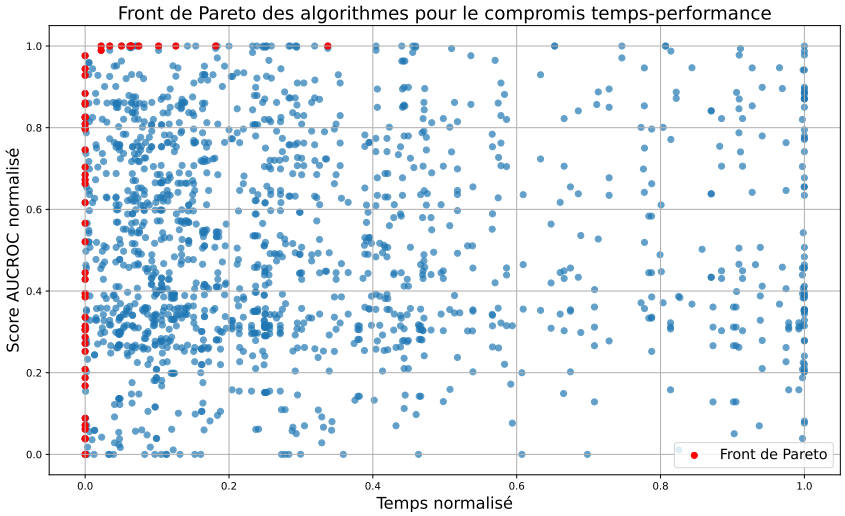
   où p_{t} et p_{p} sont les pondérations pour le temps et la performance, respectivement.

Analyse simplifiée du compromis Pour simplifier l'analyse, nous avons fusionné les résultats de chaque algorithme à travers différents corpus (Figure 5.15) :

— Chaque point sur le graphique représente la performance d'un algorithme spécifique en termes de temps de traitement moyen et de score AUCROC moyen (Figure 5.15a).
— L'étiquette de chaque point indique à quel algorithme il correspond, ce qui permet d'identifier plus facilement les algorithmes qui offrent les meilleurs compromis.
— Les points rouges sur le graphique représentent le front de Pareto, montrant les algorithmes optimaux qui équilibrent le temps et la performance.
— Le tableau 5.6 présente les algorithmes optimaux de Pareto identifiés, en affichant le temps moyen, le ROC moyen et leurs valeurs normalisées.



(a) Diagramme de dispersion des performances des algorithmes



(b) Front de Pareto

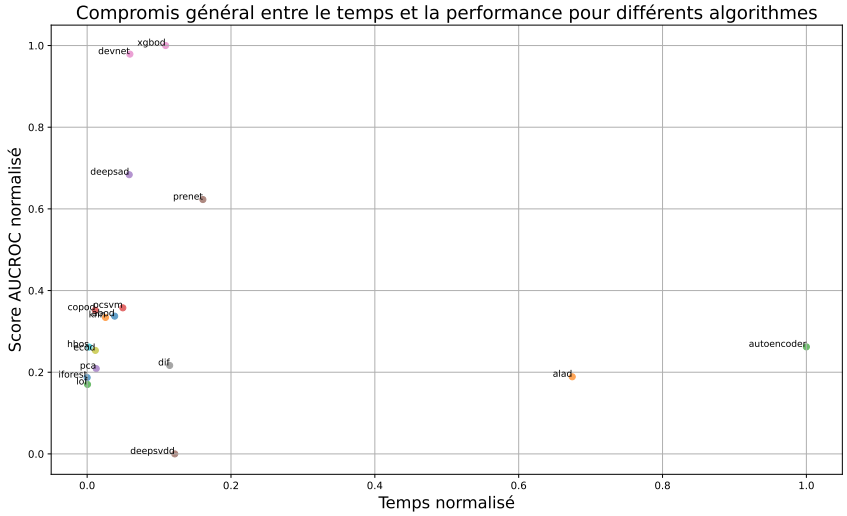
FIGURE 5.14 – Compromis entre le temps d'exécution et la performance de détection. Chaque point représente la performance d'un algorithme sur un corpus.

Sélection de l'algorithme

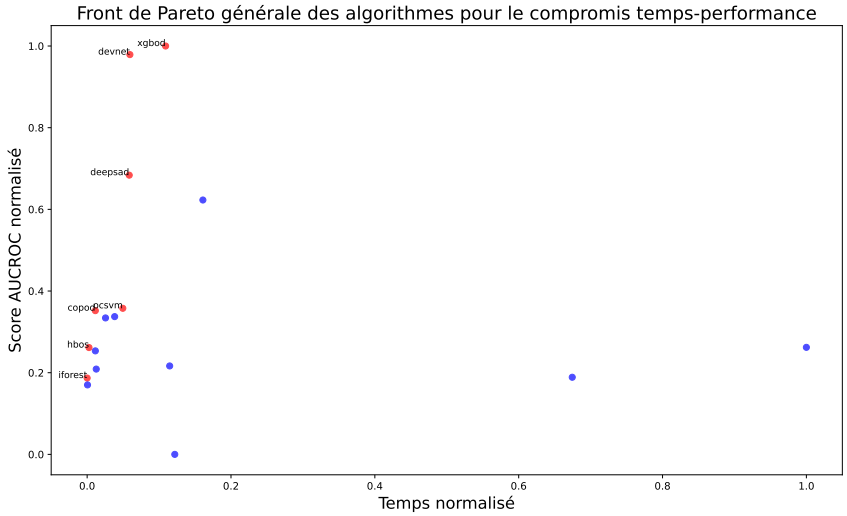
— **COPOD, HBOS, and IForest** : Ces algorithmes sont particulièrement efficaces en termes de temps et offrent de bonnes performances, ce qui les rend idéaux pour les applications en temps réel et les déploiements à grande échelle.

— **DeepSAD and DevNet** : Ces modèles DL affichent de bonnes performances avec une efficacité temporelle raisonnable, appropriés pour les tâches où la performance est critique et où les ressources sont modérément disponibles.

— **XGBOD** : Offre les meilleures performances globales, mais à un coût de calcul relativement plus élevé, convenant aux scénarios exigeant la plus grande exactitude de détection.



(a) Chaque point représente la performance d'un algorithme sur un corpus.



(b) Front de Pareto

FIGURE 5.15 – Compromis entre le temps d'exécution et la performance de détection.

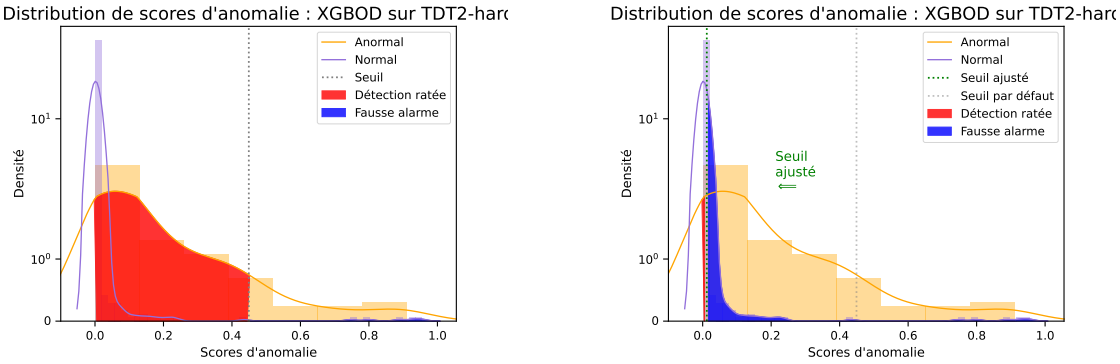
5.3.6 Seuillage

La computation des scores d'anomalie est un aspect fondamental de la recherche sur la détection d'anomalies. Cependant, un élément pratique crucial est souvent négligé : le processus de seuillage. Alors que notre travail se concentre principalement sur les modèles capables de générer des scores d'anomalie distincts pour les échantillons normaux et anormaux, l'identification d'un seuil optimal est tout aussi essentielle pour l'efficacité opérationnelle.

Des scores élevés en termes d'AUCROC et de AUCPR indiquent la capacité d'un modèle à distinguer efficacement entre les observations anormales et normales sur une gamme de seuils possibles. Ces métriques reflètent l'efficacité globale d'un mo-

Algorithme	Temps normalisé	AUCROC normalisé	Score composite
COPOD	0,0114	0,4314	0,7154
DeepSAD	0,0585	0,7297	0,8356
DevNet	0,0595	0,9644	**0,9525**
HBOS	0.0026	0,3962	0,6988
IForest	**0.0000**	0,3331	0,6665
OCSVM	0,0496	0,4621	0,7063
XGBOD	0,1091	**1,0000**	0,9455

TABLE 5.6 – Comparaison des algorithmes sélectionnés en fonction du temps normalisé, des scores AUCROC normalisés et du score composite ($p_t = 0,5$ et $p_p = 0,5$). Les résultats les meilleurs sont indiqués en gras, tandis que ceux considérés comme le deuxième et le troisième meilleurs pour chaque critère sont soulignés, afin d'identifier les choix optimaux de Pareto dans les tâches de détection d'anomalies textuelles.



(a) Performance de XGBOD avec le seuil fixé.

(b) Performance de XGBOD avec le seuil ajusté.

FIGURE 5.16 – Performance de XGBOD avant et après ajustement du seuil. Cette figure présente une comparaison des performances de XGBOD sur le corpus TDT2-hard avant et après les ajustements de seuil. La sous-figure (a) illustre les performances du modèle en utilisant le seuil initialement fixé. Elle met en évidence les difficultés d'équilibrer la précision et le rappel avec un seuil statique. La sous-figure (b) montre l'amélioration des performances après l'ajustement du seuil.

dèle pour la détection d'anomalies, mais ne garantissent pas des performances optimales à un seuil particulier.

Dans les scénarios pratiques, où un seuil statique a été initialement fixé, comme dans notre étude, la performance à ce seuil précis peut ne pas correspondre au potentiel indiqué par ces scores globaux. Pour obtenir des performances optimales ou répondre à des besoins spécifiques tels que la précision ou le rappel, il est essentiel d'ajuster le seuil en fonction du contexte d'application particulier. Un tel ajustement permet au modèle de mieux répondre aux exigences opérationnelles spécifiques, en équilibrant la sensibilité et la spécificité selon les demandes uniques de l'environnement de déploiement.

Par exemple, comme le montrent la Figure 5.16 et le Tableau 5.7, le modèle XGBOD a obtenu un score AUCROC impressionnant de 0,953 sur le corpus TDT2-hard, indiquant une forte capacité de discrimination. Cependant, avec le seuil statique-

	Seuil	Précision	Rappel	F1	Raté	Fausse alarme
Par défaut	0,4492	0,7222	0,1300	0,2203	0,8700	0,0056
Ajusté	0,0120	0,6489	0,8500	0,7359	0,1500	0,0511

TABLE 5.7 – Performance de XGBOD avant et après ajustement du seuil. Cette table détaille la performance du modèle XGBOD sur le corpus TDT2-hard avant et après les ajustements de seuil. La ligne 'par défaut' représente les configurations de seuil par défaut, montrant une performance équilibrée mais non optimale. La ligne 'ajusté' reflète les configurations du seuil ajustées, ce qui améliore considérablement le rappel et réduit le taux de ratés, bien qu'au prix d'une augmentation des fausses alarmes.

ment fixé en fonction du taux de contamination, le modèle n'a atteint qu'un F-score modeste de 0,22 avec un taux de ratés élevé de 0,87. En ajustant le seuil, nous avons amélioré le score F1 de manière substantielle à 0,736 tout en réduisant le taux de ratés à 0,15. Cet ajustement souligne l'importance du seuillage dynamique dans les applications de détection d'anomalies sur mesure.

5.4 Synthèse

Dans ce chapitre, nous avons détaillé les expériences menées pour évaluer l'efficacité et la robustesse des algorithmes de détection d'anomalies appliqués à des données textuelles. Nous avons examiné 17 algorithmes distincts sur divers corpus multilingues, couvrant trois types principaux d'anomalies textuelles : thématiques, sentimentales et discours haineux. Les analyses se concentrent sur plusieurs aspects clés : paradigmes d'apprentissage, types d'anomalies textuelles, techniques de représentation, mécanismes de calcul des scores d'anomalie, compromis temps-performance, et seuils de détection.

Les résultats montrent que les algorithmes qui introduisent un certain niveau de supervision surpassent les modèles non supervisés grâce à l'utilisation de données partiellement étiquetées. En particulier, les modèles semi-supervisés PU et faiblement supervisés améliorent la précision et le rappel, soulignant l'importance des échantillons d'anomalies annotés pour une détection efficace.

Les algorithmes de détection d'anomalies montrent une efficacité variable selon le type d'anomalie textuelle. Ils sont efficaces pour détecter les anomalies thématiques, mais leur performance diminue avec des distinctions subtiles entre sujets similaires. Les sentiments déviants posent des défis modérés, avec une variabilité de performance moindre. La détection de discours de haine s'avère plus complexe, en raison de la complexité contextuelle et linguistique des textes, nécessitant ainsi des techniques TALN plus sophistiquées pour une meilleure analyse des textes.

L'analyse a également révélé que la sélection des techniques de représentation et des mécanismes de calcul des scores d'anomalie affecte significativement la performance des algorithmes. Les représentations contextuelles avancées, comme Sentence-BERT, surpassent les méthodes traditionnelles, bien qu'elles nécessitent des ressources computationnelles accrues. Les modèles basés sur l'apprentissage du score de bout en bout, tels que DevNet et PReNET, qui optimisent directement les scores d'anomalie, montrent de meilleures performances comparées aux méthodes traditionnelles basées sur la proximité, la probabilité, et la reconstruction. Les ap-

proches d'apprentissage ensembliste, telles que XGBOD, qui combinent plusieurs détecteurs faibles, démontrent une capacité supérieure à travers différents ensembles de données.

En conclusion, nos expériences montrent que l'intégration de techniques avancées de représentation et de mécanismes de calcul des scores, ainsi que l'utilisation de données étiquetées, sont essentielles pour améliorer la détection d'anomalies textuelles dans le cadre de fouille de données. Cependant, il est crucial de choisir judicieusement les algorithmes et les approches en fonction du type de données et des exigences spécifiques, afin de maximiser l'efficacité tout en équilibrant les ressources computationnelles.

Conclusion de la deuxième partie

Dans cette partie de la thèse, nous avons cherché à combler le décalage entre les méthodologies de fouille de données et les exigences spécifiques de la détection d'anomalies dans les données textuelles. Cette exploration nous a permis de mieux comprendre comment les techniques de fouille de données peuvent être appliquées pour détecter des anomalies dans les textes.

* * *

Le chapitre 3 a posé les bases en présentant les deux éléments fondamentaux pour la détection d'anomalies dans le cadre de la fouille de données : la représentation de texte et la sélection d'algorithmes de détection appropriés. Ce chapitre a détaillé les méthodologies utilisées pour prétraiter et représenter les données textuelles, en soulignant l'importance de capturer les nuances linguistiques et contextuelles nécessaires à une détection efficace des anomalies. Il a également exploré l'adaptation de divers algorithmes de détection d'anomalies, allant des méthodes traditionnelles aux techniques incorporant des réseaux de neurones, spécifiquement conçus pour gérer la nature à haute dimensionnalité des données textuelles.

Le chapitre 4 s'est concentré sur le corpus utilisé pour nos expériences. Il a discuté des critères de sélection et d'adaptation des jeux de données textuels afin de garantir une évaluation complète des méthodes. Ce chapitre a établi une base solide en fournissant une gamme diversifiée et pertinente d'anomalies textuelles, ouvrant la voie à une validation empirique robuste des techniques de fouille de données adaptées.

Le chapitre 5 s'est appuyé sur les bases posées dans les chapitres précédents en menant une série d'expériences visant à valider empiriquement l'efficacité des méthodes de fouille de données appliquées à la détection d'anomalies textuelles. Ce chapitre a offert une analyse détaillée des performances de différents modèles selon divers points de vue et dans différents scénarios, répondant directement aux questions de recherche spécifiques posées au début de la Partie II.

* * *

Paradigmes d'apprentissage. Les expériences ont montré que les paradigmes d'apprentissage semi-supervisé et faiblement supervisé améliorent considérablement la détection d'anomalies textuelles par rapport aux approches non supervisées. L'apprentissage semi-supervisé PU (*Positive-Unlabeled*) et l'apprentissage faiblement supervisé, qui intègrent des étiquettes partiellement annotées, ont démontré une capacité supérieure à équilibrer les faux positifs et les faux négatifs. Un taux d'annotation d'environ 50% est optimal pour maximiser les performances sans rendements décroissants. Les anomalies plus marquées, lorsqu'elles sont incluses dans l'ensemble d'entraînement, se sont révélées plus efficaces pour définir des frontières de décision claires pour la détection.

Techniques de représentation. Les techniques de représentation avancées, comme Sentence-BERT (SBERT), surpassent systématiquement les méthodes traditionnelles telles que TF-IDF en matière de détection d'anomalies. Les représentations contextuelles riches fournies par SBERT capturent mieux les nuances sémantiques et contextuelles, ce qui est crucial pour détecter des anomalies subtiles dans les données textuelles. Cependant, cette amélioration de la performance s'accompagne d'une augmentation du temps de traitement et des ressources nécessaires.

Scores d'anomalie. Certains types de scores d'anomalie offrent des avantages significatifs. Les algorithmes basés sur l'apprentissage des scores d'anomalie de bout en bout, comme DevNet et PReNet, montrent des performances supérieures en présence de données partiellement annotées. Ces méthodes apprennent de manière adaptative à partir des annotations disponibles, maximisant l'efficacité même avec un nombre limité d'anomalies annotées. Les approches basées sur l'apprentissage ensembliste, comme XGBOD, bénéficient également de leur capacité à combiner plusieurs perspectives pour une détection plus robuste.

Apprentissage profond. Les modèles d'apprentissage profond surpassent souvent les modèles peu profonds dans la détection d'anomalies textuelles, en particulier dans des scénarios complexes où les anomalies sont subtiles. Les modèles d'apprentissage profond tels que DeepSAD, DevNet, et PReNet peuvent apprendre des représentations plus complexes et identifier des patterns anormaux mieux que les modèles traditionnels peu profonds. Cependant, leur efficacité est aussi liée à une demande accrue en ressources et en temps de calcul.

* * *

Dans la recherche actuelle sur la détection d'anomalies textuelles, les techniques de fouille de données jouent toujours un rôle prédominant, tandis que les techniques de TALN, en particulier les modèles de langue, ont généralement servi d'outils de soutien. Cependant, avec l'avènement et le progrès rapide des grands modèles de langue (LLMs) au cours des dernières années, le rôle des modèles de langue a évolué, passant de simples assistants à des solveurs centraux dans de nombreux domaines.

Etant donné ce tournant, la prochaine partie de cette thèse se concentrera sur l'exploration du rôle que les LLMs peuvent jouer dans la détection d'anomalies textuelles. Nous visons à étudier comment les LLMs se comportent par rapport aux méthodes traditionnelles de fouille de données et quels avantages ou nouvelles perspectives ils peuvent apporter à ce domaine. Nous déterminerons si les LLMs peuvent révolutionner la détection d'anomalies dans les textes en comparant leurs performances avec les méthodes établies, ouvrant ainsi la voie à des innovations méthodologiques dans le domaine de détection d'anomalies.

Troisième partie

Méthodes de TALN - LLMs

Introduction de la troisième partie

Comme nous l'avons vu au Chapitre 2, la détection d'anomalies textuelles a évolué de manière significative grâce aux avancées dans le domaine du traitement automatique des langues naturelles (TALN). Les modèles de langues (*Language Models*, LMs), en particulier, ont joué un rôle majeur dans cette évolution. Chaque progrès notable dans la modélisation de langue a propulsé la détection d'anomalies textuelles vers de nouveaux horizons.

Au vu des succès récents des grands modèles de langue (*Large Language Models*, LLMs) dans diverses branches de la recherche en TALN et en fouille de textes, il devient pertinent d'explorer leur potentiel spécifique dans le contexte de la détection d'anomalies textuelles. Cette troisième partie de la thèse se consacre donc à une exploration approfondie du rôle et de l'efficacité des LLMs dans ce domaine.

* * *

Avant de plonger dans le vif du sujet, un aperçu général des LLMs sera présenté au Chapitre 6. Ce chapitre établira les bases nécessaires pour comprendre comment ces modèles peuvent être exploités pour identifier des anomalies dans les textes. Par la suite, le Chapitre 7 détaillera la méthodologie adoptée pour notre étude, incluant une description des modèles sélectionnés et de l'approche de prompt utilisée. Enfin, le chapitre 8 constitue le cœur de cette étude empirique et exposera les expériences menées. Ces expériences sont conçues et organisées pour répondre aux questions de recherche suivantes :

— Compte tenu de l'efficacité prouvée des LLMs dans diverses tâches de TALN et de fouille de textes, ces modèles peuvent-ils également exceller dans la détection d'anomalies textuelles ? Si oui, comment se comparent-ils aux méthodes traditionnelles de fouille de données ?

— Quels **avantages spécifiques** ou quelles **nouvelles perspectives** les LLMs apportent-ils au domaine de la détection d'anomalies textuelles ?

— Quel **modèle** ou **type de modèle** se révèle le plus efficace pour détecter des anomalies dans les textes et dans quelles conditions ?

— Comment les différentes **techniques de conception de prompts** influencent-elles la performance des LLMs dans cette tâche ?

— Existe-t-il des types particuliers d'anomalies ou des jeux de données pour lesquels les LLMs démontrent une supériorité spécifique par rapport aux approches traditionnelles ?

— Quelles sont les **limites** de l'utilisation des LLMs dans la détection d'anomalies textuelles et comment ces défis pourraient-ils être adressés dans les recherches futures ?

A travers ces chapitres, nous visons à apporter une analyse détaillée et complète du potentiel et des limites des LLMs dans la détection d'anomalies textuelles. En répondant à ces questions, nous souhaitons fournir une compréhension plus profonde de leur capacité à transformer le domaine de la détection d'anomalies textuelles.

CHAPITRE 6

GRANDS MODÈLES DE LANGUE

Sommaire

6.1 Introduction

Dans ce chapitre, nous aborderons la base théorique de grandes modèles de langue (*Large Language Models*, LLMs) qui représentent une évolution remarquable dans le domaine du traitement automatique des langues naturelles (TALN). Ces modèles sont au cœur des progrès récents en matière de compréhension et de génération automatique de texte. Une compréhension approfondie de ces modèles est cruciale pour tirer pleinement parti de leurs capacités.

Le chapitre se structure en trois parties principales :

1. Nous débuterons par un examen rétrospectif des modèles de langues, en retraçant leur développement depuis les premiers modèles statistiques jusqu'aux récents LLMs. Cette perspective historique nous permet de comprendre les progrès technologiques et conceptuels qui ont mené aux systèmes actuels.
2. Ensuite, nous présenterons une vue d'ensemble des LLMs, en mettant l'accent sur les quatre éléments fondamentaux qui les définissent : le processus d'entraînement, l'architecture des transformeurs, la grande échelle et les capacités

émergentes. Chacun de ces aspects sera exploré pour démontrer comment ils contribuent à la performance exceptionnelle de ces modèles.

3. Enfin, nous introduirons la notion de prompt et d'ingénierie de prompts. Cette partie mettra en lumière la base théorique de la méthodologie employée dans nos expérimentations.

Ce cadre théorique offre non seulement une compréhension des fondements des LLMs mais sert également de préparer à leur application spécifique dans le domaine qui nous intéresse.

6.2 Modèles de langue

La modélisation de la langue est une technique fondamentale dans le domaine du TALN. Un modèle de langue fonctionne essentiellement comme une distribution de probabilités sur des mots ou des séquences de mots. Il est construit pour capturer les patterns de la langue humaine et prédire la probabilité d'une séquence de mots, facilitant ainsi la prédiction des mots à venir ou manquants dans un contexte spécifique. Les progrès des modèles de langues ont permis de renforcer les capacités des machines à manipuler des langues naturelles, influençant ainsi une série d'applications de TALN telles que la traduction et la synthèse automatique. L'évolution des modèles de langue peut être divisée en quatre phases clés, chacune marquant un changement substantiel de paradigme dans la manière dont les tâches de TALN sont abordées et résolues.

6.2.1 Modèles de langue statistiques

Les modèles de langue statistiques (*Statistical Language Models*, SLMs) représentent la forme la plus primitive de modélisation de langue dans le domaine du TALN. Ces modèles sont généralement basés sur l'hypothèse de Markov, qui simplifie le problème en supposant que la probabilité d'un mot ne dépend que d'un nombre fixe de mots précédents. Cela a conduit au développement de modèles `n-gramme`, où n indique le nombre de mots pris en compte pour prédire le mot suivant. Les exemples courants incluent les bigrammes (considérant un mot précédent) et les trigrammes (considérant deux mots précédents). L'idée principale derrière les SLMs est de construire un modèle prédictif pour les séquences de mots en utilisant des probabilités dérivées de données observées. Par exemple, un modèle trigramme prédit la probabilité d'un mot en fonction des deux mots précédents : $P(w_i|w_{i-1}, w_{i-2})$. Ces modèles calculent la probabilité de séquences de mots en comptant les occurrences dans un grand corpus et en utilisant ces décomptes pour estimer la probabilité de différentes séquences.

À cette époque, les modèles de langue étaient principalement utilisés pour améliorer la performance de tâches spécifiques de TALN. En fournissant un cadre statistique pour prédire les séquences de mots, les SLMs ont amélioré des tâches telles que la reconnaissance vocale, la traduction automatique et la génération de texte. La capacité à prédire le mot suivant dans une séquence permettait d'obtenir des résultats plus précis et contextuellement appropriés dans ces applications.

Le paradigme du TALN à ce stade reposait principalement sur une approche d'apprentissage supervisé, souvent décrite comme un cadre d'« `entraînement et test` ». Dans ce contexte, un modèle spécifique à une tâche était entraîné sur un jeu

de données contenant des exemples d'entrées et de sorties pertinents pour la tâche ciblée. L'accent était mis sur l'ingénierie des caractéristiques (*feature engineering*), où les experts en TALN mobilisaient leurs connaissances du domaine pour identifier et extraire les caractéristiques essentielles à partir des textes bruts. Ces caractéristiques servaient ensuite à entraîner des modèles dotés du biais inductif approprié pour apprendre efficacement à partir des données disponibles.

Globalement, les modèles de langue statistiques ont jeté les bases des avancées ultérieures dans le domaine du TALN, en démontrant l'importance des approches statistiques pour comprendre et générer la langue humaine. L'accent mis sur l'ingénierie des caractéristiques et l'apprentissage supervisé à cette étape a fourni des idées et des techniques précieuses qui ont influencé les développements ultérieurs dans le domaine.

6.2.2 Modèles de langue neuronaux

L'évolution des modèles de langue a connu un bond significatif dans les années 2010 avec l'essor des modèles de langues neuronaux (*Neural Language Models*, NLMs). Cette période a marqué un départ des approches statistiques traditionnelles, en inaugurant une ère où les réseaux neuronaux ont commencé à modéliser la probabilité des séquences de mots.

L'une des principales avancées de cette époque a été la mise en pratique de représentations distribuées des mots, également connues sous le nom de « *word embeddings* » (littéralement « plongements de mots »). Les plongements de mots encodent les mots sous forme de vecteurs denses dans un espace vectoriel continu, capturant les relations sémantiques plus efficacement que les modèles précédents. L'introduction de Word2Vec [Mikolov et al., 2013], un modèle qui utilise un réseau neuronal à deux couches pour apprendre ces représentations de mots distribuées, a constitué une avancée décisive à cet égard. Il utilisait deux approches principales : le *Continuous Bag of Words* (CBOW) et le Skip-Gram. Le CBOW prédit un mot en fonction de son contexte environnant, tandis que le Skip-Gram prédit les mots du contexte environnant à partir d'un mot cible. Ces modèles ont démontré que les plongements de mots pouvaient capturer efficacement les relations syntaxiques et sémantiques dans le texte, ce qui s'est avéré très utile pour diverses tâches de TALN.

Le paradigme du TALN à ce stade adhérait encore largement au cadre d'« `entraînement et test` », mais a évolué vers une solution de bout en bout (*end-to-end*) plus sophistiquée. Les modèles de langue neuronaux sont devenus une partie intégrante des solutions de bout en bout pour les tâches de TALN, servant souvent de couche de plongements dans des réseaux neuronaux plus étendus. Cette intégration a transformé les modèles de langue neuronaux en apprenants de caractéristiques adaptés à la tâche. Ces modèles ont appris des représentations riches des mots et des phrases, permettant des applications de TALN plus sophistiquées et contextuellement conscientes. Cette évolution a marqué la transition du TALN de l'ingénierie manuelle des caractéristiques à l'ingénierie de l'architecture, où la conception d'architectures de réseau appropriées est devenue primordiale.

Des modèles tels que Word2Vec [Mikolov et al., 2013] et GloVe [Pennington et al., 2014] ont joué un rôle déterminant à cette époque, en influençant considérablement le développement d'architectures neuronales plus complexes et en inspirant d'autres recherches sur les plongements de mots et leurs applications. Ces avancées ont mis

en évidence la capacité des modèles de langue neuronaux à capturer des patterns linguistiques complexes et à améliorer les performances des tâches de TALN.

En résumé, cette phase de l'évolution des modèles de langue a introduit des réseaux neuronaux qui ont fondamentalement changé le TALN en permettant un traitement de la langue plus nuancé et plus conscient du contexte. Elle a marqué un tournant majeur par rapport aux méthodes statistiques antérieures, ouvrant la voie aux innovations ultérieures en matière d'architectures neuronales et de modèles pré-entraînés qui conduiraient à de nouveaux progrès dans le domaine du TALN.

6.2.3 Modèles de langue pré-entraînés

La période de 2017 à 2019 a marqué une ère décisive dans l'évolution des modèles de langue avec l'avènement des modèles de langue pré-entraînés (*Pre-trained Language Models*, PLMs). Cette phase a bouleversé le paysage du TALN en introduisant un nouveau paradigme : le « `pré-entraînement et ajustement` (*pre-train and fine-tuning*) » . Les modèles de langue pré-entraînés ont apporté un changement profond dans la façon d'aborder les tâches de TALN, en positionnant les modèles de langue comme des solveurs de problèmes transférables plutôt que des outils spécifiques à une tâche.

Les PLMs impliquent deux phases principales : le pré-entraînement et l'ajustement.

— **Pré-entraînement** : Cette phase initiale consiste à entraîner le modèle pour qu'il apprenne à prédire la probabilité des séquences de texte observées. Le pré-entraînement utilise de vastes quantités de données textuelles brutes disponibles sur Internet, ce qui permet au modèle d'apprendre des caractéristiques riches et polyvalentes de la langue. Ce processus repose sur un apprentissage non supervisé ou auto-supervisé, au cours duquel le modèle capte des patterns syntaxiques et sémantiques ainsi que des connaissances factuelles à partir des données, en optimisant des objectifs de modélisation tels que la prédiction de mots masqués ou de la phrase suivante dans une séquence.

— **Ajustement** : Après le pré-entraînement, le modèle peut est affiné pour des tâches spécifiques (appelées tâches en aval) par un entraînement supplémentaire sur un jeu de données plus restreint et spécifique à la tâche. L'ajustement consiste à introduire des paramètres supplémentaires et à ajuster le modèle pré-entraîné en fonction d'objectifs spécifiques à la tâche, tels que l'analyse de sentiments, la réponse à des questions ou la synthèse de texte. Cette phase permet d'adapter la compréhension générale de la langue acquise pendant le pré-entraînement aux exigences particulières de chaque tâche, offrant ainsi au modèle pré-entraîné des performances élevées dans un large éventail de tâches de TALN avec un minimum de données et d'efforts supplémentaires.

L'introduction des modèles de langue pré-entraînés a recentré la recherche en TALN sur l'ingénierie des objectifs, qui consiste à concevoir des objectifs d'entraînement optimaux pour les phases de pré-entraînement et d'ajustement Liu et al. [2021b]. Les chercheurs s'efforcent d'affiner ces objectifs pour améliorer la capacité des modèles à se généraliser efficacement à travers diverses tâches et domaines. Ce nouveau paradigme a conduit au développement de modèles pré-entraînés une seule fois, puis ajustés à plusieurs reprises pour différentes tâches, réduisant ainsi de ma-

nière significative les ressources computationnelles et le temps nécessaires par rapport à l'entraînement de modèles distincts pour chaque tâche.

Parmi les modèles les plus remarquables de cette époque, on trouve BERT, GPT-2 et BART. BERT (*Bidirectional Encoder Representations from Transformers*) [Devlin et al., 2019] s'appuie sur une architecture de transformeurs bidirectionnels pour effectuer un pré-entraînement sur un objectif de modélisation de langue masquée et de prédiction de la phrase suivante. Ces deux objectifs permettent à BERT de saisir des informations contextuelles dans les deux sens, ce qui améliore considérablement ses performances dans des tâches nécessitant une compréhension globale du texte. GPT-2 (*Generative Pre-trained Transformer 2*) [Radford et al., 2019], quant à lui, utilise un transfomeur unidirectionnel pour générer des textes cohérents et contextuellement pertinents, démontrant ainsi la puissance des modèles de langue pour la génération de texte. BART (*Bidirectional and Auto-Regressive Transformers*) [Lewis et al., 2020] combine les approches de BERT et GPT en utilisant des transfomeurs à la fois bidirectionnels et autorégressifs, permettant d'exceller dans les tâches de génération et de compréhension de textes.

Cette période a marqué l'essor des modèles de langue comme des outils polyvalents et puissants pour le TALN, capables de transférer des connaissances à travers différentes tâches et domaines. Le paradigme de « `pré-entraînement et ajustement` » a permis de créer des modèles dotés de capacités inégalées en compréhension et génération du langage, établissant un nouveau standard de performance et d'efficacité en TALN.

6.2.4 Grands modèles de langue

La dernière évolution des modèles de langue, amorcée autour de 2020, est marquée par l'émergence des grands modèles de langue (*Large Language Models*, LLMs). Ces modèles représentent une avancée considérable dans la modélisation de la langue, grâce à plusieurs facteurs clés : l'architecture des transformeurs, la disponibilité massive de données d'entraînement, l'augmentation des capacités de calcul (GPUs, TPUs, clusters) et les stratégies d'entraînement distribuées. Ces avancées ont collectivement permis le développement de modèles d'une ampleur et d'une capacité sans précédent.

Les LLMs ont provoqué un changement de paradigme significatif en TALN avec l'introduction du cadre « `pré-entraînement, prompt, prédiction` » [Liu et al., 2021b]. Contrairement à l'époque précédente qui reposait fortement sur l'ajustement de modèles pré-entraînés pour des tâches spécifiques, cette nouvelle approche exploite les capacités et les connaissances inhérentes des LLMs. Plus précisément, au lieu d'adapter des modèles pré-entraînés à chaque tâche en aval par un ajustement complexe, les tâches sont maintenant reformulées pour s'aligner sur l'entraînement original du modèle à l'aide de prompts soigneusement conçus.

Partie centrale du paradigme « `pré-entraînement, prompt, prédiction` », les prompts sont des instructions textuelles qui guident le modèle pour produire le résultat souhaité. Par exemple, pour une tâche de traduction avec un LLM, un prompt pourrait être : « Traduisez la phrase suivante de l'anglais vers le français : '*Hello, how are you ?*' ». Le modèle, grâce à ses vastes connaissances pré-entraînées, exécute la demande et génère la traduction : « Bonjour, comment allez-vous ? ». Cette approche exploite les connaissances déjà acquises par le modèle et ses capacités polyvalentes,

lui permettant d'accomplir des tâches spécifiques sans nécessiter un entraînement supplémentaire. Des prompts bien formulés jouent un rôle central dans ce contexte, car ils exploitent pleinement les connaissances pré-acquises du modèle pour guider efficacement ses réponses.

L'ingénierie de prompt est ainsi devenue un chaînon clé dans l'utilisation des LLMs. Elle consiste à concevoir des prompts clairs et concis, capables de susciter des réponses correctes de la part du modèle. Le défi réside dans la création de prompts qui s'alignent avec l'entraînement du modèle tout en guidant efficacement ses prédictions pour accomplir les tâches visées.

À ce stade, le rôle des modèles de langue a évolué vers celui d'un solutionneur de problèmes généraux. Le vaste pré-entraînement sur des jeux de données diversifiés leur permet d'aborder diverses tâches sans ajustements significatifs spécifiques à la tâche. Les LLMs sont ainsi très adaptables et efficaces, capables de générer des textes cohérents et contextuellement pertinents, de répondre à des questions, de traduire des textes et même d'effectuer des tâches d'écriture créative, en fonction des prompts qu'ils reçoivent. Parmi les modèles notables de cette époque, on trouve GPT d'OpenAI, LlaMA de Meta, et Claude d'Anthropic. Ces modèles ont établi de nouvelles normes en matière de compréhension et de génération de langue.

Le paradigme « `pré-entraînement, prompt, prédiction` » a fondamentalement transformé le paysage du TALN. Non seulement il améliore l'efficacité et la polyvalence des modèles de langue, mais il démocratise également l'accès aux techniques avancées de TALN. Désormais, grâce à des prompts simples, même les non-experts peuvent tirer parti de la puissance des LLMs, rendant les applications TALN sophistiquées plus accessibles.

6.3 Grands modèles de langue

6.3.1 Aperçu général des LLMs

Les LLMs constituent la frontière la plus avancée dans l'évolution de la modélisation de la langue, symbolisant une rupture majeure par rapport aux approches antérieures. Après l'ère des modèles statistiques et la transition vers les modèles neuronaux et pré-entraînés, les LLMs inaugurent une nouvelle phase où la taille et la complexité des modèles jouent un rôle déterminant.

Ces modèles avancés se distinguent par quatre éléments fondamentaux qui définissent leur performance et leur fonctionnement : le **processus d'entraînement plus élaboré**, l'**architecture des transformeurs**, la **grande échelle** et les **capacités émergentes**. Dans les sections suivantes, nous explorerons chacun de ces aspects pour comprendre comment ils permettent aux LLMs d'atteindre une efficacité sans précédent et de transformer le domaine du TALN.

6.3.2 Entraînement

Les LLMs représentent une évolution avancée des PLMs, avec un processus d'entraînement plus complexe qui introduit des distinctions significatives par rapport aux PLMs traditionnels tels que BERT, GPT-2, ELECTRA et ALBERT. Le processus général d'entraînement des LLMs comporte plusieurs étapes clés : pré-entraînement, ajustement et RLHF.

Pré-entraînement Le pré-entraînement reste une étape fondamentale dans le développement des LLMs, où le modèle est exposé à de vastes ensembles de données non étiquetées pour apprendre les structures et les patterns de la langue.

Contrairement aux PLMs traditionnels comme BERT et ELECTRA, qui utilisent divers objectifs d'entraînement tels que la modélisation de langue masquée ou la détection de tokens remplacés, la majorité des LLMs privilégient la modélisation de langue autorégressive comme stratégie de pré-entraînement. Dans cette approche, le modèle génère chaque token en fonction des tokens précédents, ce qui le rend particulièrement efficace pour des tâches comme la génération de texte, les systèmes de dialogue, et la création de contenu.

Étant donnée une séquence de tokens $\mathbf{x}_{1:T} = [x_1, x_2, \ldots, x_T]$, la probabilité conjointe $\mathbb{P}(x_{1:T})$ est modélisée comme un produit de probabilités conditionnelles :

$$\mathbb{P}(x_{1:T}) = \prod_{t=1}^{T} \mathbb{P}(x_t|\mathbf{x}_{0:t-1})$$

où x_0 est un token spécial représentant le début de la séquence. Chaque probabilité conditionnelle $\mathbb{P}(x_t|\mathbf{x}_{0:t-1})$ est calculée par le modèle, où les tokens précédents $\mathbf{x}_{0:t-1}$, autrement dit le contexte, sont encodés en représentations de haute dimension par l'encodeur du modèle $f_{enc}(\cdot)$. La probabilité conditionnelle finale est ensuite calculée comme suit :

$$\mathbb{P}(x_t|\mathbf{x}_{0:t-1}) = g_{LM}(f_{enc}(\mathbf{x}_{0:t-1}))$$

où g_{LM} est la couche de prédiction du modèle, qui mappe le contexte encodé à une distribution de probabilité sur le vocabulaire, permettant de prédire le prochain token de la séquence.

Cet entraînement autorégressif est optimisé à l'aide de l'estimation du maximum de vraisemblance, où le modèle apprend à maximiser la vraisemblance des séquences de tokens observées dans le corpus d'entraînement, améliorant ainsi sa capacité à prédire les tokens à venir en fonction du contexte passé.

Ajustement Après la phase de pré-entraînement, les LLMs subissent généralement une phase d'ajustement pour s'adapter à des tâches spécifiques ou aux besoins des utilisateurs. Cette étape implique souvent un apprentissage supervisé, où le modèle est entraîné sur des ensembles de données étiquetées plus restreints, conçus pour des tâches particulières comme la traduction, le résumé ou la réponse à des questions.

Une avancée notable des LLMs est l'ajustement par instruction (*instruction tuning*). Contrairement aux PLMs classiques comme BERT ou T5, qui nécessitaient un ajustement spécifique à chaque tâche, des modèles comme GPT-4 et LLaMA 3 sont ajustés à l'aide de grands ensembles de données d'instructions ou de prompts effectués par des humains couvrant de multiples tâches. Ce processus permet au modèle de mieux se généraliser à de nouvelles tâches et de bien performer même dans des scénarios d'apprentissage de *zero-shot* ou de *few-shot*. L'ajustement par instruction améliore spécifiquement la capacité du modèle à suivre des commandes en langue naturelle, le rendant plus polyvalent et capable de traiter une grande variété de tâches sans nécessiter un entraînement spécialisé pour chacune d'elles.

RLHF Bien que l'ajustement par instruction aide à aligner les modèles sur les commandes humaines, l'apprentissage par renforcement avec feedback humain (*Reinforcement Learning from Human Feedback*, RLHF) pousse cet alignement encore plus loin en affinant activement les réponses du modèle sur la base des évaluations humaines. Dans le cadre du RLHF, des évaluateurs humains fournissent des retours sur la qualité, la pertinence et l'utilité des réponses du modèle. Ces retours sont intégrés dans une boucle d'apprentissage par renforcement, permettant au modèle de s'améliorer en générant des réponses plus utiles et mieux alignées avec les attentes humaines.

Alors que l'ajustement par instruction se concentre sur la généralisation des tâches et la compréhension des commandes, le RLHF garantit que les réponses du modèle sont optimisées en fonction des préférences humaines, telles que la clarté, l'utilité et le ton. Cela est particulièrement important dans les applications interactives, comme les agents conversationnels ou les assistants virtuels, où la génération de réponses conformes aux attentes des utilisateurs est essentielle. Le RLHF apporte une couche supplémentaire de raffinement aux LLMs, les aidant non seulement à produire des réponses précises, mais aussi à répondre de manière plus intuitive et alignée sur les besoins des utilisateurs.

6.3.3 Transformeur et auto-attention

Le transformeur (*transformer*), proposé par Vaswani et al. [2017], a marqué une avancée majeure dans le domaine du TALN en introduisant une nouvelle architecture qui excelle dans la manipulation des dépendances à longue distance dans le texte. Contrairement aux réseaux neuronaux récurrents ou convolutionnels traditionnels, le transformeur repose sur un mécanisme d'auto-attention pour traiter les séquences de texte, ce qui lui permet de prendre en compte le contexte global de chaque mot en relation avec tous les autres mots de la séquence de manière simultanée. Cette capacité à gérer efficacement le contexte textuel permet une meilleure compréhension et génération de texte, ce qui fait des transformeurs un pilier central dans la conception des LLMs.

Mécanisme d'auto-attention Le mécanisme d'attention dans les transformeurs permet au modèle de se concentrer sur différentes parties de la séquence d'entrée lors de la génération de chaque mot en sortie. Au lieu de traiter les séquences de manière strictement séquentielle, l'attention permet au modèle de pondérer l'importance de chaque mot en fonction de sa relation avec les autres mots de la séquence. Cela permet de gérer plus efficacement les dépendances à longue distance.

L'auto-attention (*self-attention*), une variante spécifique de l'attention, permet au modèle de capturer les relations entre tous les mots d'une séquence en parallèle, ce qui améliore considérablement la compréhension du contexte global. Pour chaque mot d'une séquence, trois vecteurs sont calculés : la requête (*query*), la clé (*key*) et la valeur (*value*). Le modèle calcule les scores d'attention en prenant le produit scalaire de la requête avec toutes les clés de la séquence, puis divise ces scores par la racine carrée de la dimension des clés pour stabiliser l'apprentissage. Ces scores sont ensuite normalisés à l'aide d'une fonction softmax pour obtenir les poids d'attention, qui sont utilisés pour pondérer les valeurs correspondantes. Cette somme pondérée devient la nouvelle représentation du mot dans la séquence.

L'auto-attention permet également une parallélisation efficace. Contrairement aux réseaux récurrents, qui traitent les tokens séquentiellement, l'auto-attention permet de traiter toutes les positions de la séquence simultanément, grâce à des opérations matricielles, rendant ainsi possible un traitement massivement parallèle sur des architectures matérielles modernes comme les GPU. Cette capacité à traiter les séquences en parallèle rend l'architecture du transformeur particulièrement adaptée à l'entraînement sur de grands ensembles de données.

L'attention multi-tête (*multi-head attention*) est une extension du mécanisme d'auto-attention. Elle consiste à exécuter plusieurs opérations d'attention en parallèle, avec chaque tête se concentrant sur des aspects différents des relations entre les mots. Les sorties de toutes les têtes d'attention sont ensuite concaténées et transformées linéairement pour produire la représentation finale. Cette approche permet au modèle de capturer une diversité de dépendances et de patterns dans le texte, améliorant ainsi la compréhension des relations complexes entre les mots d'une séquence.

De récentes innovations, telles que l'attention éparse (*sparse attention*), permettent de traiter des contextes encore plus longs de manière plus efficace en réduisant les coûts de calcul. Modèles comme Gemini exploitent cette approche pour gérer des contextes étendus, essentiels dans des tâches complexes comme la détection d'anomalies textuelles. Ces avancées renforcent la capacité des LLMs à analyser de grandes quantités de données tout en conservant des performances élevées.

Architecture de transformeur L'architecture de base du transformeur est composée d'un encodeur et d'un décodeur, chacun constitué de plusieurs couches. L'encodeur prend en charge le traitement du texte d'entrée pour produire des représentations contextuelles riches, tandis que le décodeur utilise ces représentations pour générer le texte de sortie. Toutefois, dans les LLMs recents, l'architecture est souvent simplifiée en utilisant uniquement le décodeur, une configuration particulièrement bien adaptée aux tâches de génération de texte.

Les modèles comme GPT-4 et LlaMA 3 illustrent parfaitement cette tendance, où le décodeur reçoit une séquence d'entrée et produit un texte fluide et cohérent en s'appuyant sur les tokens précédents. Contrairement aux architectures plus complexes qui incluent un encodeur pour la compréhension du texte, les modèles à base de décodeurs se concentrent exclusivement sur la production de texte en se basant sur l'historique des tokens déjà générés, ce qui permet d'exceller dans des tâches comme la création de dialogues ou la complétion de phrases.

En outre, cette configuration permet un entraînement plus rapide en évitant le double traitement (encodeur et décodeur). Les modèles décodeurs sont également hautement adaptables, se prêtant bien aux tâches non supervisées, où ils peuvent être pré-entraînés sur de vastes corpus de données textuelles non étiquetées grâce à la modélisation de langue autorégressifs.

6.3.4 La grande échelle

La grande échelle est la caractéristique la plus distinctive des LLMs, qui leur permet d'exécuter des tâches complexes et de saisir les subtilités de la langue. Le concept de « grande échelle » englobe trois facteurs essentiels, comme indiqué par Kaplan et al. [2020] : la taille du modèle, qui fait référence au nombre de paramètres au sein des réseaux de neurones ; la taille des données d'entraînement, soit le nombre

total de tokens dans le corpus d'entraînement ; et le coût computationnel, mesuré en termes d'opérations en virgule flottante par seconde (*Floating-point operations per second*, FLOPs). De nombreuses études ont démontré que l'augmentation de ces facteurs améliore considérablement les capacités des LLMs [Radford et al., 2019; Brown et al., 2020; Chowdhery et al., 2022].

Le développement de lois d'échelle, telles que la loi d'échelle Kaplan-McCandlish (KM) [Kaplan et al., 2020] et la loi d'échelle Chinchilla [Hoffmann et al., 2022], a fourni des cadres quantitatifs qui guident l'optimisation de la taille du modèle, du volume des données et des ressources computationnelles. La loi KM suggère que l'augmentation de la taille du modèle a un impact plus significatif que celle de la taille des données d'entraînement pour améliorer les performances. Cette étude recommande d'entraîner de très grands modèles sur une quantité relativement modeste de données, en arrêtant l'entraînement bien avant d'atteindre la convergence. À l'inverse, la loi d'échelle Chinchilla prône une approche équilibrée, où la taille du modèle et le volume des données d'entraînement doivent être ajustés proportionnellement pour maximiser à la fois l'efficacité et les performances.

Ces lois d'échelle non seulement améliorent notre compréhension de la conception des LLMs, mais elles affinent également les stratégies d'entraînement, permettant de créer des modèles à la fois performants et computationnellement efficaces.

6.3.5 Capacités émergentes et apprentissage en contexte

À mesure que l'échelle des modèles augmente, ils manifestent souvent des capacités sophistiquées qui ne sont pas observées dans les modèles plus petits. Communément appelées « capacités émergentes » [Wei et al., 2022b], ces capacités englobent un large éventail de fonctionnalités, allant de tâches simples comme les opérations arithmétiques et l'exécution de code à des activités complexes telles que la planification et la manipulation d'outils. Ces capacités ne sont pas explicitement programmées, mais plutôt apparaissent naturellement à la suite d'un entraînement extensif sur des ensembles de données diversifiés.

L'émergence de ces capacités ne peuvent pas être prédites simplement en extrapolant les performances de modèles plus petits. En revanche, elles se matérialisent sous forme d'un « saut discontinu » dans les performances, analogue à une transition de phase en physique [Wei et al., 2022b]. Ce bond significatif se produit généralement lorsque les modèles atteignent une certaine échelle, leur permettant d'acquérir une compréhension avancée de la langue et d'autres compétences cognitives qui n'étaient ni prévues ni observées chez des modèles plus petits. Cette progression non linéaire souligne la dynamique complexe qui accompagne l'évolution des LLMs et révèle leur potentiel à débloquer de nouveaux niveaux de performance cognitive et analytique.

Parmi les capacités émergentes des LLMs, les plus remarquables sont l'apprentissage en contexte (*in-context learning*), le suivi des instructions et le raisonnement par étapes :

— **Apprentissage en contexte** : Les LLMs excellent en tant qu'apprenants de *few-shot*, capables de s'adapter à de nouvelles tâches avec une exposition minimale, en utilisant seulement quelques exemples d'entrée-sortie pour guider leur comportement [Brown et al., 2020]. Cette capacité, qui imite les processus d'apprentissage humains, montre leur aptitude à généraliser efficacement à partir de données limitées pour aborder de nouvelles tâches.

— **Suivi des instructions** : Grâce à leur flexibilité, les LLMs peuvent comprendre et exécuter des tâches entièrement nouvelles en se basant uniquement sur des instructions descriptives, sans nécessiter de réajustement spécifique [Wei et al., 2022a]. Cette capacité une application directe dans divers domaines.

— **Raisonnement par étapes** : Les LLMs sont capables de suivre une séquence d'étapes logiques pour arriver à une conclusion, ce qui facilite une prise de décision plus transparente et interprétable [Wei et al., 2023]. Cette capacité est crucial dans les scénarios de résolution de problèmes complexes, où comprendre le raisonnement du modèle est aussi important que la réponse elle-même.

Ensemble, ces capacités émergentes transforment les LLMs en solutionneurs de problèmes polyvalents et autonomes, capables de relever divers défis sous le paradigme de l'apprentissage à base de prompt.

6.4 Prompt

6.4.1 Prompt et apprentissage à base de prompt

Dans les systèmes informatiques, un prompt, ou une invite, désign un message textuel qui indique que le système est prêt à recevoir une entrée et à exécuter une commande. À l'instar de cette notion, dans les systèmes d'intelligence artificielle, en particulier en TALN, un prompt est un message textuel qui décrit la tâche à effectuer et guide le comportement du modèle.

Avec l'avènement des LLMs, notamment ceux dotés des capacités émergentes telles que l'apprentissage en contexte et le suivi des instructions, un changement de paradigme s'est produit, passant de l'approche traditionnelle « `pré-entraînement et ajustement` » à « `pré-entraînement, prompt et prédiction` » [Liu et al., 2021b]. Ce nouveau paradigme exploite la capacité du modèle à apprendre avec peu (*few-shot*) ou pas d'exemples (*zero-shot*). Dans ce contexte, au lieu d'ajuster les modèles pré-entraînés sur des tâches spécifiques en aval, ces tâches sont reformulées pour ressembler aux problèmes rencontrés lors de la phase d'entraînement initial. Concrètement, la tâche est présentée au modèle sous la forme d'un prompt textuel, transformant ainsi la tâche en un problème de prédiction de prochain token, un domaine où les modèles génératifs autorégressifs excellent.

Le noyau de ce nouveau paradigme réside dans l'ingénierie de prompt, c'est-à-dire la méthode systématique et pratique de conception de prompts efficaces.

6.4.2 Ingénierie de prompt

L'ingénierie de prompt consiste à concevoir une fonction de prompt $f_{prompt}(x)$ qui génère des prompts permettant la performance la plus efficace sur la tâche en aval [Qiu et al., 2020]. Ce processus comprend généralement deux étapes principales :

1. **Optimisation de la composition et de la structure** : Cette étape implique la composition et la structuration délibérées des prompts afin de maximiser la probabilité d'obtenir la réponse souhaitée de la part du modèle. Elle nécessite souvent des tests itératifs, une analyse détaillée des erreurs et des raffinements.
2. **Création de templates** : Il s'agit de concevoir des templates ou des cadres adaptables pour les prompts qui peuvent être systématiquement appliqués à diverses tâches ou requêtes afin d'assurer cohérence et efficacité.

Conception des composants Dans sa forme la plus simple, un prompt peut consister en une instruction de tâche pour l'apprentissage de *zero-shot* ou une démonstration de tâche avec quelques exemples d'entrée-sortie pour l'apprentissage de *few-shot*. Bien que ces approches soient efficaces pour de nombreuses tâches, elles sont souvent insuffisantes pour traiter des défis plus complexes. Pour améliorer la performance sur de telles tâches, diverses techniques ont été développées, notamment le prompt par chaîne de pensée (*Chain of Thought*, CoT), le prompt par arbre de pensée (*Tree of Thought*, ToT), le prompt par auto-consistance *(Self-Consistency*) et le prompt par auto-affinage (*Self-Refinement*). Chacune de ces techniques est conçue pour guider le processus de raisonnement du modèle de manière plus efficace et aligner ses sorties plus étroitement sur les patterns de raisonnement humain.

Conception des templates Un template est un cadre structuré qui guide la création de prompt. Il comprend typiquement des espaces réservés pour les données d'entrée, des instructions pour diriger le comportement du modèle ou des directives pour s'assurer que le prompt suscite des réponses appropriées. Les templates peuvent être développés manuellement ou automatiquement, ces derniers utilisant souvent des techniques avancées telles que la fouille de prompt (*prompt mining*), la paraphrase de prompt, la recherche prompt basée sur le gradient (*gradient-based prompt search*) et la notation de prompt (*prompt scoring*). Ces méthodes facilitent la découverte et l'optimisation de prmompts efficaces, améliorant ainsi la performance du modèle sur un large éventail de tâches.

6.5 Synthèse

Dans ce chapitre, nous avons exploré l'impact des modèles de langue sur le TALN. Nous avons commencé par un aperçu de l'évolution des modèles de langue, en passant des approches statistiques, comme les modèles n-grammes, aux modèles neuronaux tels que Word2Vec et GloVe, puis aux modèles pré-entraînés comme BERT et GPT. Ce parcours historique a montré comment chaque étape a apporté des améliorations dans la compréhension et le traitement du texte.

Le cœur du chapitre s'est concentré sur les LLMs, caractérisés par leur grande échelle, leur architecture à base de transformeurs, ainsi que leur capacité émergente. Nous avons discuté de l'importance du pré-entraînement et de l'ajustement, en mettant l'accent sur les modèles autorégressifs. De plus, nous avons examiné le rôle du prompting, en soulignant comment des prompts bien conçus peuvent orienter les LLMs dans des applications spécifiques sans nécessiter un ré-entraînement intensif.

En conclusion, les LLMs représentent un saut transformationnel dans le TALN, combinant une puissance de calcul considérable à des paradigms d'apprentissage complexes. Ces modèles sont capables de réaliser un large éventail de tâches avec une précision remarquable, établissant un nouveau standard pour l'analyse et la génération de texte. Leur capacité à se généraliser à travers diverses tâches, guidée par l'ingénierie des prompts, offre un potentiel immense pour des applications avancées telles que la détection d'anomalies textuelles, que nous explorerons dans les chapitres suivants.

CHAPITRE 7

MÉTHODOLOGIE

Sommaire

7.1 Introduction

Après avoir exploré les fondements théoriques des grands modèles de langue, notamment leurs capacités émergentes, dans le chapitre précédent, nous nous concentrons ici sur leur application pratique à la détection d'anomalies textuelles. Ce chapitre expose en détail la méthodologie employée pour exploiter pleinement le potentiel de ces modèles.

Nous commençons par une analyse des principales familles de LLMs, telles que GPT, LlaMA, Mistral et Gemini, afin d'identifier les modèles les plus adaptés à notre problématique. Bien que tous ces modèles soient basés sur l'architecture des transformeurs, ils se distinguent par leur taille, leur conception et leurs processus d'entraînement, ce qui offre une diversité d'options pour la détection d'anomalies textuelles.

Le cœur de notre méthodologie repose sur l'utilisation des LLMs en tant que solutionneurs de problèmes, où l'art du prompting joue un rôle central. Nous décrivons les composants essentiels des prompts ainsi que les stratégies de construction et d'optimisation, visant à maximiser la performance des modèles pour la détection des écarts textuels. Enfin, nous abordons l'importance de la segmentation des corpus, indispensable pour permettre aux LLMs de traiter des ensembles de données volumineux de manière efficace.

En combinant ces approches, ce chapitre présente une méthodologie complète pour adapter les LLMs à des tâches spécifiques, tout en assurant un haut niveau de performance dans la détection d'anomalies textuelles.

7.2 Modèles

Récemment, le domaine du TALN a connu des avancées majeures grâce au développement d'une variété de LLMs par des entités commerciales et des organisations académiques. Ces modèles diffèrent largement en termes d'architecture, de capacités et d'accessibilité. Afin de mener une investigation concentrée et approfondie sur la détection d'anomalies textuelles à l'aide des LLMs, nous avons réalisé une recherche préparatoire pour déterminer les modèles à privilégier. Dans ce cadre, nous avons étudié un large éventail de modèles, en nous basant sur l'examen des documentations officielles et les évaluations issues de plusieurs *leaderboards* de LLMs, notamment LMSYS, Artificial Analysis, et Leaderboard LLM de HuggingFace. Après cette étude préparatoire, nous avons décidé de nous concentrer sur quatre familles de modèles : GPT, LlaMA, Mistral, et Gemini, couvrant à la fois des modèles open-source et propriétaires.

Dans cette section, nous présenterons une introduction générale de ces modèles, en mettant en lumière leur architecture, leurs processus d'entraînement, et leurs données d'entraînement. Cependant, il convient de noter que, pour certains modèles, notamment les modèles propriétaires, les détails techniques restent non divulgués par leurs développeurs. Notre discussion se conformera strictement aux informations publiées, évitant toute spéculation ou estimation par les chercheurs afin de garantir l'intégrité et la véracité de notre analyse.

7.2.1 Les modèles de la famille GPT

La série *Generative Pre-trained Transformer* (GPT)[1], introduite par OpenAI, constitue une pierre angulaire dans l'évolution des LLMs. Depuis le dévoilement du GPT original par Radford et al. [2018], cette famille de modèles n'a cessé de s'étendre et d'évoluer, repoussant les limites de ce que l'intelligence artificielle peut accomplir en matière de compréhension et de génération de la langue naturelle. Les modèles de la famille GPT sont construits sur une architecture de transformeur autoregressive à décodeur unique, qui excelle dans la génération de textes cohérents et contextuellement pertinents en prédisant le mot suivant dans une séquence basée sur les mots précédents. Cette conception, ainsi que le pré-entraînement sur des ensembles de données étendus de textes non étiquetés, leur permet de générer des contenus textuels semblables à ceux produits par les humains dans une vaste gamme de tâches. Depuis l'introduction du premier modèle GPT en 2018, la série a évolué à travers plusieurs générations, chacun étant plus capable et sophistiqué que le précédent.

GPT-1 [Radford et al., 2018], qui a marqué le début de cette série innovante, est un modèle pionnier exploitant l'architecture des transformeurs en se concentrant uniquement sur la partie décodeur. Ce modèle comprend douze couches de transformeurs, chacune avec douze têtes d'auto-attention, où chaque tête comporte des états de 64 dimensions, totalisant ainsi 768 dimensions par couche. Le modèle, riche de 117 millions de paramètres, utilise l'algorithme d'optimisation Adam avec

1. Dans cet thèse, le terme « GPT » est utilisé de manière restreinte pour désigner la famille de modèles proposée par OpenAI. Toutefois, dans un sens plus large, le terme désigne un type de LLM prédominant dans l'intelligence artificielle générative. Les GPTs sont basés sur l'architecture transformeur, pré-entraînés sur de grands ensembles de données textuelles non étiquetées, et capables de générer du contenu nouveau semblable à celui produit par les humains. Bien que cette notion ait été proposée par OpenAI et que les modèles les plus célèbres soient développés par cette entreprise, il existe également d'autres modèles GPT tels que Neo-GPT.

un taux d'apprentissage maximal de 2.5e-4, augmenté linéairement puis diminué selon un calendrier cosinus. Pour l'activation, il emploie l'unité linéaire d'erreur gaussienne (GELU). Le modèle suit un processus d'entraînement en deux étapes : un pré-entraînement non supervisé suivi d'un ajustement spécifique à la tâche. Le pré-entraînement est réalisé sur le BooksCorpus, un ensemble de données de 4.5 GB provenant de 7000 livres de fiction non publiés, ce qui permet au modèle de gérer efficacement les informations à longue portée.

GPT-2 [Radford et al., 2019] a considérablement développé cette base, non seulement en taille, avec des options allant jusqu'à 1,5 milliard de paramètres, mais aussi en portée, en utilisant un ensemble de données beaucoup plus large dérivé de millions de pages web. Ce modèle a démontré des améliorations remarquables dans la génération de textes plus longs et plus complexes, montrant le potentiel de l'augmentation de la taille des modèles de transformeurs.

GPT-3 [Brown et al., 2020] a marqué une avancée majeure dans la série avec ses 175 milliards de paramètres. Entraîné sur un corpus de texte encore plus vaste, il a introduit la capacité d'apprentissage de *few-shot* et même de *zero-shot*, lui permettant d'exécuter des tâches sans nécessiter d'exemples explicites. Les capacités de GPT-3 ont établi de nouveaux standards pour les modèles de langage, tant en termes de polyvalence que de profondeur des connaissances. L'évolution de cette série s'est poursuivie avec **GPT-3.5** ainsi que ses versions ajustées, telles que **ChatGPT** et **InstructGPT**, conçues pour des tâches plus interactives et instructives, augmentant ainsi l'utilité des modèles dans des applications d'intelligence artificielle plus complexes.

GPT-4, sorti en mars 2023, a étendu ces avancées en introduisant des capacités multimodales, permettant au modèle de traiter non seulement du texte mais aussi des images, et en offrant des réponses plus nuancées et conscientes du contexte. L'architecture de GPT-4 comportait également des innovations comme le message du système, ce qui permet des sorties plus contrôlées et adaptées aux instructions spécifiques des utilisateurs. Plus récemment, **GPT-4o**, lancé en 2024, a marqué une avancée significative en intégrant des capacités à générer des sorties à travers plusieurs modalités, y compris le texte, l'audio et les images en temps réel.

7.2.2 Les modèles de la famille LlaMA

La famille de modèles LlaMA (*Large Language Model Meta AI*) est une série de LLMs développée par Meta AI. À l'instar des modèles GPT, les LLaMA sont des modèles autorégressifs basés sur des transformeurs à décodeur unique. Cependant, ils intègrent plusieurs modifications uniques visant à améliorer les performances et la stabilité. Ces innovations comprennent la normalisation des couches par la méthode RMS (*Root Mean Squared*), les fonctions d'activation SwiGLU, et les plongements RoPE (*Rotary Positional Embeddings*). Lancée en février 2023, la série LLaMA a évolué de manière significative à chaque nouvelle version, renforçant son efficacité globale et sa fiabilité par rapport à des modèles comme GPT.

La version initiale, **LlaMA 1** [Touvron et al., 2023a], offre des configurations allant de 6,7 milliards à 65,2 milliards de paramètres. Ces modèles ont été entraînés sur un vaste ensemble de données comprenant entre 1 et 1,4 trillion de tokens issus de sources diversifiées telles que CommonCrawl, GitHub, et Wikipedia. LlaMA 1 est conçu comme un modèle fondamental généraliste, adaptable à diverses d'applications

spécifiques. **LlaMA 2** [Touvron et al., 2023b] a augmenté les capacités des modèles précédents jusqu'à 69 milliards de paramètres, avec un corpus d'entraînement élargi à 2 trillions de tokens. LLaMA 2 propose à la fois des versions fondamentales et des versions ajustées pour les instructions, qui bénéficient d'améliorations notables en termes d'alignement et de performance, grâce à l'intégration du RLHF.

Le modèle **LLaMA 3**, lancée en avril 2024, constitue la troisième génération de la série, repoussant encore plus loin les limites avec des modèles contenant jusqu'à 70 milliards de paramètres et utilisant un ensemble de données étendu à 15 trillions de tokens. Ce modèle est également disponible en deux versions : une version fondamentale et une version ajustée pour les instructions. LlaMA 3 bénéficie d'importantes optimisations architecturales, incluant un vocabulaire étendu de 128 000 tokens et un tokenizer amélioré, qui permettent un encodage de la langue plus efficace. L'application de l'attention par requêtes groupées améliore également l'efficacité de l'inférence. De plus, la taille de la fenêtre de contexte a été étendue à 8192 tokens, ce qui permet d'aborder des tâches complexes comme la fouille de textes.

Le modèle a été entraîné sur plus de 15 trillions de tokens provenant de sources publiques, avec une attention particulière portée à la qualité des données, grâce à des techniques avancées de filtrage, incluant des filtres heuristiques et la détection de contenu NSFW (*Not Safe For Work*). Cette rigueur garantit un entraînement sur des données de haute qualité, élément crucial pour la robustesse du modèle. Le jeu de données d'entraînement est principalement composé de données en anglais. Toutefois, pour préparer le modèle à des applications multilingues, plus de 5% des données de pré-entraînement proviennent de plus de 30 autres langues. Malgré cette inclusion, les performances du modèle dans ces langues ne sont pas censées égaler celles obtenues en anglais.

7.2.3 Les modèles de la famille Mistral

Les modèles Mistral, développés par la société française Mistral AI, constituent une gamme diversifiée de LLMs, incluant des versions open source ainsi que des modèles propriétaires optimisés.

Mistral 7B, lancé en septembre 2023, est le premier modèle de cette famille. À l'instar des LLMs précédents comme les GPT et les LlaMA, Mistral 7B est un modèle autorégressive à base de transformeur à décodeur unique. Il utilise des mécanismes d'attention groupée et d'attention à fenêtre coulissante (*Sliding Window Attention*, SWA) pour améliorer les performances et gérer des séquences plus longues. Avec uniquement 7,3 milliards de paramètres, ce modèle est déclaré plus performants en anglais et en code que les modèles LlaMA 2 13B et LlaMA 1 34B.

Mixtral 8×7B, sorti en décembre 2023, est le premier modèle de Mistral à base d'une architecture de mélange clairsemé d'experts (*Sparse Mixture-of-Experts*, SMoE). Dans cette architecture, les couches traditionnelles de réseaux de neurones denses sont remplacées par des couches d'experts. Les experts sont des réseaux de neurones spécialisés, chacun traitant une partie spécifique des données d'entrée. Le modèle comporte 8 groupes d'experts, chacun avec ses propres paramètres distincts. À chaque couche et pour chaque token, un réseau de routage détermine quels experts traiteront le token, en sélectionnant deux groupes d'experts dont les sorties sont combinées de manière additive. Ainsi, bien que le Mistral 8×7B comporte un total de 46,7 milliards de paramètres, chaque token n'en utilise activement que 12,9

milliards. Cette configuration permet au modèle de maintenir l'efficacité computationnelle d'un modèle plus petit, tout en bénéficiant la puissance d'un large espace de paramètres. Le Mixtral 8×7B a démontré des performances exceptionnelles dans divers benchmarks, surpassant même des modèles plus grands tels que LlaMA 70B et GPT-3.5.

Enfin, le modèle **Mistral 8×22B**, lancé en avril 2024, est actuellement le modèle open source le plus performant de Mistral. Il utilise 39 milliards de paramètres actifs sur un total de 141 milliards, offrant une efficacité inégalée pour sa taille. Ce modèle dispose d'une fenêtre de contexte de 64 000 tokens, permettant une récupération précise de l'information à partir de grands documents. Ce modèle est optimisé pour suivre des instructions précises et est fluide en plusieurs langues, dont l'anglais, le français, l'italien, l'allemand et l'espagnol.

7.2.4 Les modèles de la famille Gemini

La famille de modèles Gemini est une série de LLMs multimodaux développée par Google DeepMind. Annoncée en décembre 2023, cette série a été conçue comme une alternative compétitive au modèle GPT-4 d'OpenAI.

Les modèles de première génération, **Gemini 1.0** [Team et al., 2024], se déclinent en trois variantes : Ultra, Pro et Nano. Comme les autres LLMs, ils reposent sur une architecture de transformateur à décodeur unique, mais avec des modifications pour permettre un entraînement et une inférence efficaces sur les unités de traitement de tenseurs (TPUs) de Google. **Gemini Ultra** se distingue par ses capacités à exceller dans des tâches complexes, surpassant GPT-4 et d'autres modèles concurrents sur divers benchmarks industriels. **Gemini Pro** est optimisé pour offrir un bon compromis entre performance, coût et latence, tandis que **Gemini Nano** est conçu pour les dispositifs à mémoire limitée.

La deuxième génération, **Gemini 1.5 Pro** [Team et al., 2024], introduit des innovations substantielles, avec deux aspects particulièrement notables. Premièrement, l'implémentation de l'architecture de mélange clairsemé d'experts, similaire à celle des modèles Mistral, améliore l'efficacité sans sacrifier les performances. Deuxièmement, une avancée majeure réside dans la gestion des contextes longs. La taille de la fenêtre de contexte a été étendue à plusieurs millions de tokens, ce qui rend Gemini 1.5 incontournable pour les tâches impliquant la compréhension de contextes longs, notamment la détection d'anomalies textuelles. Enfin, **Gemini 1.5 Flash**, la version distillée du Pro, conserve une longueur de contexte supérieure à 2 millions de tokens, tout en étant plus léger et efficace.

7.3 LLMs en tant que solutionneur

Comme discuté dans §6.2, avec l'avènement des LLMs, le rôle des modèles de langue a évolué, passant de simples assistants à des fournisseurs de solutions autonomes pour une variété de tâches. Cette section explore la méthodologie pour employer les LLMs en tant que solutionneurs indépendants spécifiquement pour le problème de détection d'anomalies textuelles.

L'utilisation de LLMs, tels que GPT-4 et LlaMA 3, comme fournisseurs de solutions repose principalement sur une approche basée sur les prompts. Il s'agit d'inter-

agir avec le modèle à l'aide des prompts textuels soigneusement conçus qui guident le modèle pour générer des complétions de texte appropriées. Comme les modèles sont particulièrement sensibles aux nuances de la conception des prompts, il est essentiel de maîtriser l'art de construire des prompts efficaces, ce que l'on appelle l'ingénierie des prompts. Ainsi, la conception et la construction des prompts constituent le cœur de notre méthodologie.

Dans la discussion qui suit, nous examinerons les composants essentiels des prompts efficaces, les stratégies pour leur construction et le processus de conversion du corpus original en entrées appropriées pour les LLMs.

7.3.1 Composants de prompt

La construction de prompts pour les LLMs peut commencer simplement, soit avec une instruction pour un prompt *zero-shot*, soit en fournissant quelques exemples pour un prompt *few-shot*. Le prompt *zero-shot* ne nécessite aucun exemple préalable et consiste à présenter au modèle une instruction ou une requête directe Wei et al. [2022a]. Par exemple, un prompt *zero-shot* pour la traduction pourrait se présenter ainsi :

Prompt *zero-shot* pour la traduction

Traduisez le texte suivant de l'anglais en français :
"[Texte à traduire]" ⇒

En revanche, le prompt *few-shot*, comme démontré par Brown et al. [2020], consiste à fournir au modèle un petit ensemble d'exemples pour l'aider à comprendre la tâche ciblée et la sortie désirée :

Prompt *three-shot* pour l'nalyse de sentiments

Texte : "J'adore le nouveau design du site web ! Il est tellement convivial et attrayant."
Sentiment : Positif

Texte : "Le service au restaurant était terrible. J'ai attendu une heure avant que ma nourriture n'arrive."
Sentiment : Négatif

Texte : "L'événement était bien organisé, mais le lieu était trop petit pour le nombre de participants."
Sentiment : Neutre

Texte : "[Texte à analyser]"
Sentiment :

Ces formes de base de prompt peuvent servir efficacement de nombreuses tâches, en particulier celles impliquant un raisonnement simple ou pour lesquelles le modèle a été spécifiquement ajusté. Cependant, les nouvelles tâches complexes exigent souvent des prompts plus élaborés pour obtenir les résultats souhaités.

Dans notre recherche, nous avons adopté une structure de prompt enrichie et étendue comprenant plusieurs éléments clés : le message du système, les instructions, les exemples, la chaîne de pensée, et les indications de sortie. Cette conception nous permet d'adapter les prompts pour répondre aux exigences spécifiques à

la tâche de détection d'anomalies textuelles, améliorant ainsi les performances du modèle pour produire des sorties pertinentes et exactes.

7.3.1.1 Message du système

Un message du système est un composant essentiel placé au tout début d'un prompt lors de l'interaction avec des LLMs. Ce message textuel définit le cadre de l'interaction en fournissant au modèle le contexte nécessaire, les instructions ou les informations pertinentes pour la tâche ou le scénario ciblé. Ce message préliminaire est ainsi également appelé méta-prompt, prompt du système ou message d'initialisation.

Largement utilisés depuis la publication du GPT-3, les messages du système sont devenus un concept formalisé et une pratique recommandée avec le GPT-4, et ont été adoptés par des modèles ultérieurs tels que LLaMA 2 et Mistral. Le but de ces messages consiste à améliorer la guidabilité (*steerability*) du modèle, en d'autres termes, sa capacité à aligner son comportement et ses réponses sur les intentions humaines et des objectifs spécifiques, évitant ainsi les résultats non désirés.

Le texte du message du système est traité avec une attention particulière par le modèle, ayant un impact plus significatif et plus global sur les réponses du modèle par rapport au message de l'utilisateur ou d'autres éléments contextuels dans le prompt. Des techniques avancées, telles que *Ghost Attention*, ont été développées pour garantir que le message du système influence de manière cohérente le comportement du modèle tout au long de l'interaction [Touvron et al., 2023a].

En termes pratiques, le message du système définit le profile du modèle, ses capacités et les restrictions en fonction d'un scénario spécifique, ce qui permet de délimiter son cadre opérationnel. Sans ces directives précises, les LLMs, qui sont extrêmement polyvalents, pourraient jouer par défaut le rôle d'un assistant IA générique, susceptible de répondre à un large éventail de problèmes sans discernement.

Par exemple, dans la détection d'anomalies textuelles, un prompt générique pourrait conduire le modèle à signaler toute forme d'irrégularité textuelle, des fautes d'orthographe aux incohérences stylistiques. Cependant, un message du système bien défini peut diriger le modèle pour qu'il se concentre sur des anomalies spécifiques pertinentes à un rôle ou à un scénario particulier, comme les erreurs grammaticales pour un enseignant de français ou les événements exceptionnels pour un analyste de renseignements. Si nous définissons plus précisément le scénario d'application comme étant celui de la veille concurrentielle et stratégique, le modèle devient capable de filtrer les textes non pertinents, en se concentrant exclusivement sur ceux relatifs aux contextes économiques et technologiques.

Pour la détection d'anomalies textuelles, nous avons conçu et comparé deux types de messages du système :

Message du système générique

- **Rôle** : Spécialiste en fouille de textes, avec un intérêt particulier pour la détection d'anomalies.
- **Capacité** : Identifier les déviations significatives dans les grands ensembles de textes.
- **Exemple** :

Message du système générique

You are a text mining specialist with expertise in anomaly detection. Your skills enable you to identify significant deviations in content within large text datasets.

Message du système spécifique au scénario

— **Rôle** : Analyste de renseignements spécialisé dans la détection de textes déviants.
— **Capacité** : Détecter des déviations significatives dans le contenu textuel.
— **Scénario** : Scénarios tels que la veille médiatique, la surveillance de la réputation en ligne et la cybersécurité.
— **Exemple** :

Message du système spécifique au scénario

You are an intelligence analyst specializing in detecting anomalous texts in large datasets. Your expertise enable you to identify significant deviations in media content, providing insights during media monitoring.

L'utilisation appropriée du message du système permet de diriger l'IA à opérer avec des paramètres étroitement définis, ce qui améliore la pertinence et la fiabilité de ses réponses pour des applications ciblées.

7.3.1.2 Instruction

L'instruction constitue le composant le plus fondamental d'un prompt, souvent utilisé pour guider le modèle sur la tâche attendue. Pour les tâches simples et conventionnelles, un prompt *zero-shot* avec des instructions directes peut être suffisant. Ces instructions, généralement dérivées directement de la définition de la tâche, indiquent clairement au modèle les actions à entreprendre.

En pratique, les instructions sont généralement utilisées conjointement avec le message du système. Pour des tâches simples, elles peuvent être directement intégrées dans ce message. En revanche, pour des tâches plus complexes, elles servent de complément en fournissant des détails opérationnels supplémentaires que le message système ne couvre pas entièrement.

Dans notre étude, il est pertinent de développer deux ensembles d'instructions correspondant à nos deux types de messages du système. Plus précisément, pour les messages du système centrés sur le scénario, les instructions pourraient préciser des actions telles que « identifier les sentiments déviants », « reconnaître les thématiques aberrantes » ou « détecter les discours nuisibles ».

Cependant, cette stratégie soulève certaines préoccupations. . Les LLMs utilisés dans cette étude ont été ajustés pour des tâches comme la classification des thématiques et l'analyse des sentiments, ou ont subi un apprentissage par renforcement avec retour humain pour filtrer les discours haineux. Fournir des instructions détaillées pour des anomalies spécifiques, telles que les sentiments déviants ou les thématiques aberrantes, pourrait involontairement amener les modèles à s'appuyer sur leurs connaissances et biais préexistants, interprétant ainsi la détection d'anomalies comme une simple extension de tâches qu'ils maîtrisent déjà. Cela risquerait de

fausser la tâche de détection, où le modèle semblerait améliorer ses performances, non pas en détectant les anomalies de manière plus efficace, mais en appliquant à tort son expérience dans des domaines connexes.

Nos expériences préliminaires sur la construction de prompts ont confirmé ces préoccupations, notamment dans la détection des sentiments déviants et du langage nuisible. Lorsque des instructions détaillées sont fournies, les modèles ont tendance à aborder ces tâches comme de simples analyses de sentiment ou détections de discours haineux avec une distribution des étiquettes extrêmement déséquilibrée. Ils exploitent ainsi des capacités développées pour d'autres tâches, plutôt que d'établir une norme et d'identifier de véritables anomalies.

Par conséquent, dans notre étude, nous avons décidé de limiter les instructions détaillées uniquement à la détection d'anomalies thématiques, où nous pouvons restreindre les informations complémentaires pour éviter ces écueils. Pour les autres types d'anomalies, nous éviterons de fournir des normes explicites ou des définitions spécifiques de ce qui constitue une anomalie. Nos instructions adhéreront étroitement à une définition générique de la détection d'anomalies textuelles, en se concentrant uniquement sur l'aspect opérationnel de l'identification des écarts significatifs, sans spécifier la nature de ces déviations.

Les instructions génériques et spécifiques pour les anomalies thématiques sont basées sur les définitions suivantes :

Définition générique

Une anomalie textuelle est une instance qui, dans un contexte donné, présente une déviation significative par rapport à la majorité des textes.

Définition spécifique au scénario

Dans un corpus de veille médiatique, un texte anormal est défini comme un texte rapportant des événements ou des incidents qui s'écartent de manière significative par rapport à ceux décrits dans la majorité des autres textes du corpus.

Cette approche permet aux LLMs d'utiliser leurs capacités à établir, en toute indépendance, des normes à partir de la majorité des textes et à identifier des déviations sans notions préconçues sur ce qui constitue une anomalie. De cette manière, nous visons à véritablement tester et améliorer la capacité innée des modèles à détecter de véritables anomalies textuelles.

7.3.1.3 Exemples

Bien que les LLMs montrent des capacités impressionnantes en *zero-shot*, ils rencontrent souvent des difficultés face à des tâches plus complexes sans exemples préalables. Pour y remédier, le prompt *few-shot* est employé, ce qui permet au modèle d'apprendre de manière contextuelle en intégrant des exemples spécifiques dans le prompt. Cette approche aide à orienter le modèle vers de meilleures performances.

Dans les scénarios d'apprentissage *few-shot*, les exemples illustratifs, particulièrement ceux tirés du jeu de données d'entraînement, jouent un rôle crucial. Notre

étude se focalise sur deux catégories principales d'exemples, chacune adaptée à différents paradigmes d'apprentissage :

1. **Exemples *Normal-Only*** : Ces exemples sont composés uniquement d'instances normales (négatives). Cela s'aligne sur le paradigme d'apprentissage semi-supervisé NO, où le modèle est entraîné exclusivement sur des instances négatives.
2. **Exemples Contaminés** : Ces exemples contiennent un mélange d'instances normales et anormales, adaptés aux paradigmes d'apprentissage semi-supervisé PU (*Positive-Unlabeled*) et faiblement supervisé.

A titre d'illustration, le contenu de ces exemples est présenté comme suit :

Exemples *Normal-Only*

Examples
Here are some normal examples to help you learn the norm : [text 1] *Asia's financial crisis deepened today. Hong Kong's stock market fell nearly 9% ...* | | *normal* (0)
...
[text n] *The U.S. dollar edged lower against the yen. On Thursday, the Nihon Keizai newspaper said ...* | | *normal* (0)

Expected Output

Exemples contaminés

Examples
[text 1] *Asia's financial crisis deepened today. Hong Kong's stock market fell nearly 9% ...* | | *normal* (0)
[text 2] *Hundreds of Indonesians have been killed in rioting over the past week, triggered by the country's economic crisis ...* | | *anomaly* (1)
...
[text n] *The U.S. dollar edged lower against the yen. On Thursday, the Nihon Keizai newspaper said ...* | | *normal* (0)

Expected Output

Pour la conception des exemples, plusieurs considérations clés sont prises en compte :

— **Étiquettes** : L'espace d'étiquetage dans les démonstrations est important [Min et al., 2022a]. Les étiquettes aident à définir le contexte de chaque exemple, améliorant ainsi la performance du modèle même avec une attribution aléatoire des étiquettes. Nous avons intégré des étiquettes dans les deux types de démonstrations. Pour les exemples contaminés, nous avons exploré un scénario d'étiquetage faible, où chaque texte est étiqueté, et un scénario PU, où seuls les exemples anormaux sont étiquetés, laissant les exemples normaux sans étiquette.
— **Sélection d'échantillons** : Nous visons à sélectionner des exemples qui ne sont ni trop facilement distinguables, ni indiscernables. Les échantillons sont choisis en fonction de leurs scores d'anomalie obtenus via des détecteurs d'anomalies non supervisés. Nous privilégions ceux ayant des scores médians ou situés dans un intervalle autour du score médian de leur groupe respectif.

— **Taille de l'ensemble de démonstration** : La distribution des étiquettes dans cet ensemble de démonstration peut également influencer les résultats. Nous avons testé différentes combinaisons, telles que des ratios de $1-4$ (anomalie à normal), $1-9$ (qui reflète la distribution réelle), $2-5$, $2-8$, et $2-18$, afin de identifier l'équilibre le plus efficace pour l'entraînement.

Ces choix dans la conception des exemples ont pour but d'améliorer la compréhension de la tâche par le modèle, lui permettant ainsi de détecter les anomalies dans les textes en interprétant efficacement les exemples présentés.

7.3.1.4 Chaîne de pensée

Le prompt *few-shot* a démontré son efficacité dans une variété d'applications grâce à la fourniture d'exemples de guidage, il montre pourtant ses limites dans les tâches nécessitant un raisonnement complexe ou une résolution de problèmes en plusieurs étapes. Pour combler cette lacune, le prompt de chaîne de pensée (*Chain of Thought*, CoT) est employé.

L'approche CoT s'inspire de la manière dont les humains abordent les problèmes complexes. Face à une tâche difficile, ils la décomposent en étapes plus petites et plus gérables, et traitent chaque partie jusqu'à la résolution du problème global. Cette approche structurée assure une progression claire et logique, minimise les risques d'erreurs et augmente l'efficacité de la résolution. En appliquant le prompt CoT, nous pouvons améliorer les capacités de raisonnement des LLMs, en les guidant à traiter les problèmes à travers une séquence d'étapes intermédiaires, similaire aux processus cognitifs humains.

Dans le contexte de la détection d'anomalies textuelles, nous utilisons le CoT en décomposant la tâche en trois étapes principales :

Décomposition de la tâche en 3 étapes

1. Établir une norme ou un standard basé sur les patterns courants observés dans la majorité des textes donnés.
2. Calculer la déviation de chaque texte par rapport à la norme établie.
3. Classer chaque texte comme normal ou anormal, selon la déviation calculée.

Cette décomposition de la tâche complexe de détection d'anomalies en sous-tâches spécifiques transforme le défi en une série de tâches plus simples et connexes. Ces tâches sont similaires à des activités telles que la synthèse de documents multiples, la synthèse de texte et la classification de texte, domaines où les LLMs ont déjà démontré leur compétence.

Les avantages de l'approche CoT dépassent l'amélioration des performances du modèle. L'un de ses principaux atouts est l'interprétabilité supplémentaire qu'elle offre. Dans des domaines tels que la détection d'anomalies, notamment avec des données textuelles, il est crucial de comprendre comment les décisions sont prises. Le prompt CoT clarifie le processus de raisonnement du modèle, expliquant comment les conclusions sont tirées. Cette transparence permet non seulement de valider le processus, mais aussi d'identifier et de corriger d'éventuelles erreurs, ce qui est essentiel pour améliorer l'efficacité et la fiabilité des applications de détection d'anomalies utilisant des LLMs.

7.3.1.5 Indication de sortie

L'élément final de la conception de notre prompt concerne la spécification de la sortie, qui détermine le format des réponses du modèle. La conception de cet indicateur influe directement sur la nature et la qualité des réponses du modèle. En demandant au modèle de fournir explicitement la « norme » établie dans l'analyse, ainsi qu'une explication pour chaque anomalie détectée, nous veillons à ce que le modèle adhère aux instructions de la CoT.

Comme abordé dans le Chapitre 1, les sorties d'un système de détection d'anomalies peuvent typiquement prendre deux formes : un score d'anomalie continu ou une étiquette binaire. Lors de nos tests préliminaires pour la construction de prompts, nous avons exploré les deux formats de sortie. Cependant, nous avons observé que, lorsqu'ils devaient générer un score d'anomalie continu, les LLMs avaient tendance à produire ces scores de manière arbitraire. Bien que la distribution des scores soit cohérente au sein d'une même requête, permettant un classement correct des textes selon leur niveau d'anomalie, les scores variaient considérablement entre les différentes requêtes, rendant les comparaisons peu fiables. Cette incohérence montre que les LLMs ont des difficultés à générer des scores continus fiables et consistants.

Face à ces défis, nous avons opté pour des étiquettes binaires dans la conception de notre prompt, simplifiant ainsi la sortie en catégorisant chaque texte comme « normal » (0) ou « anormalous » (1). Nous avons également configuré le modèle pour qu'il formate ses sorties en JSON, ce qui améliore la structure et l'accessibilité des données. Chaque sortie comprend :

— La norme établie, dérivée de l'analyse des textes normaux.

— Une prédiction binaire pour chaque texte.

— Une explication pour chaque anomalie détectée.

Ci-dessous un exemple illustrant la sortie :

Exemple de la sortie au format JSON

```
{
    "norm": "[Details of the established norm.]",
    "texts": [
        {
            "text_id": ,
            "prediction": 0
        },
        {
            "text_id": ,
            "prediction": 1,
            "explanation": "[This text deviates from the
               norm due to X reason.]"
        }
    ]
}
```

7.3.2 Construction de templates

Une fois les composants de notre prompt déterminés, l'étape suivante consiste à construire concrètement les templates de prompt, dans lesquels nos données d'entrée peuvent être intégrées de manière fluide.

L'élaboration du style et de la structure optimaux des prompts a été réalisée à partir d'essais sur un échantillon de 100 documents issus des corpus TDT2 et Amazon (en). Ce processus de conception a été facilité par l'assistant LLM GPT-4o, qui a aidé à affiner la formulation et l'organisation des prompts.

Chaque version du prompt élaboré a ensuite été évaluée sur les deux corpus, en utilisant deux LLMs tiers : WizardLM-2-8×22B et Claude 3 Sonnet, sélectionnés pour éviter tout biais en faveur de certains modèles ciblés lors de notre évaluation systématique. Pour chaque composant du prompt, nous avons exploré trois à quatre variations, en ajustant le style et la structure selon les retours du GPT-4o. La version avec le meilleur F-score a été retenue comme template final.

Durant nos tests, nous avons analysé les réponses générées pour identifier et rectifier les problèmes, notamment :

— **Spécification de la sortie** : Constatant que les modèles produisaient parfois des descriptions de norme trop exhaustives, nous avons limité la longueur de ces normes à 20 tokens et précisé les instructions telles que « Assurez-vous que cette norme capture les similitudes substantielles présentes dans les textes. ». Cet ajustement vise à concentrer le modèle sur la synthèse des patterns communs plutôt que sur l'énumération de tous les détails.

— **Instructions** : Nous avons également affiné les instructions pour assurer que le modèle limite ses évaluations aux textes fournis, en précisant explicitement : « Limitez vos évaluations aux textes fournis, sans tirer parti de contextes ou de connaissances externes. » Cette directive était nécessaire pour éviter que le modèle n'utilise ses connaissances du monde réel préexistantes, ce qui pourrait introduire des biais, notamment dans l'établissement de la norme. La norme doit être exclusivement déterminée en fonction du contexte ou du corpus analysé, sans référence à des connaissances extérieures. Ce biais, introduit par les connaissances préexistantes du modèle, pourrait également influencer son jugement lors de l'évaluation des écarts, compromettant ainsi la fiabilité de la détection des anomalies.

— **Réorganisation structurelle** : Nous avons également observé un biais de récence chez les modèles, où les instructions à la fin du prompt étaient plus susceptibles d'être suivies que celles au milieu. Par exemple, la directive « Attention : STRICTEMENT fournir UNIQUEMENT la sortie en JSON . . . » initialement placée dans la section de l'instruction de la sortie, était souvent négligée. Cependant, une fois déplacée vers la section des exemples à la fin du prompt, la conformité s'est améliorée.

Ci-dessous, un exemple de notre template de prompt final :

Exemple du template de prompt

System Message
You are a text mining specialist with expertise in text anomaly detection. Your skills enable you to identify significant deviations in content within large text datasets.
Instructions
Your task is to analyze a set of n texts and identify any text that deviates significantly from the majority in terms of content.
CoT
Step by step Instructions
Step 1. Read and Establish Norms : Carefully read each text in the collection. Identify and summarize the common patterns shared by the majority of the texts to establish a clear and concise norm. Ensure this norm captures the substantial commonalities present in the texts.
Step 2. Evaluate and Identify Deviations Evaluate each text by identifying and measuring deviations from the established norm. Keep your assessments confined to the provided texts, without leveraging external contexts or knowledge.
Step 3. Classify and Explain : Classify each text as "normal" or "anomalous" based on the identified deviations. A text is "anomalous" if it significantly deviates from the norm. Provide clear and detailed explanations for the identified anomalies, explicitly referencing the specific deviations from the norm.
Output Specification
Output Specification
In the response, STRICTLY adhere to the following JSON format :
"""json
...
"""
Examples
Examples
[text 1] Asia's financial crisis deepened today. Hong Kong's stock market fell nearly 9% ... || normal (0)
[text 2] Hundreds of Indonesians have been killed in rioting over the past week, triggered by the country's economic crisis ... || anomaly (1)
...
[text n] The U.S. dollar edged lower against the yen. On Thursday, the Nihon Keizai newspaper said ... || normal (0)
Expected Output
...
Attention :
STRICTLY provide ONLY the JSON output. DO NOT INCLUDE any introductory text, explanations, or analysis outside the structured JSON response.
Input
Texts to analyse : ["Entrée : textes à analyser"]

7.3.3 Entrées - segmentation de corpus

La détection d'anomalies textuelles se distingue des tâches traditionnelles de fouille de textes, telles que la classification ou l'analyse des sentiments, par sa forte dépendance au contexte plutôt qu'à des critères prédéfinis ou des connaissances préalables. Contrairement à ces tâches, où des modèles pré-entraînés peuvent s'appuyer sur des catégories fixes (thèmes, sentiments, etc.), la détection d'anomalies repose uniquement sur l'analyse du corpus, sans recours à des informations externes.

L'anomalie n'est pas une caractéristique intrinsèque d'un texte, mais se définit en relation avec la norme établie par l'ensemble des textes du corpus. Cette norme découle de la distribution des données dans le corpus et n'est pas déterminée a priori. Le modèle doit donc inférer la normalité à partir du corpus pour identifier les anomalies, ce qui rend l'analyse contextuelle cruciale. Ce qui est perçu comme normal dans un corpus peut être considéré comme anormal dans un autre.

Traditionnellement, le contexte étendu nécessaire à la détection d'anomalies textuelles représentait un défi majeur, en raison des limitations des fenêtres de contexte des LLMs. Les modèles antérieurs ne pouvaient traiter que de petits extraits de texte, ce qui rendait difficile l'établissement d'une norme complète à partir de grands ensembles de données.

Pour contourner ce problème, nous avons adopté une stratégie de segmentation heuristique. Cette méthode consiste à diviser le corpus en segments plus gérables, chaque segment comprenant plusieurs textes, traités séquentiellement par le modèle. Chaque segment permet au modèle de développer une norme basée sur le contexte local, évitant ainsi les difficultés liées à l'analyse d'un contexte global en une seule passe. Le modèle analyse chaque segment individuellement et prend des décisions en fonction de ce contexte localisé.

Nous avons expérimenté différentes configurations de fenêtres de segmentation, avec des tailles variant de 10 à 100 textes par segment, selon les capacités du modèle. Chaque texte est limité à 150 tokens, ce qui garantit que le modèle dispose de suffisamment d'informations pour formuler des décisions éclairées, tout en évitant une surcharge de données.

Néanmoins, il est important de noter que les progrès récents, notamment avec des modèles comme Gemini, ont considérablement amélioré la capacité des LLMs à traiter de plus grands contextes. Avec des fenêtres de contexte atteignant jusqu'à un million de tokens, ces avancées permettent d'intégrer des segments beaucoup plus larges, facilitant ainsi l'identification des anomalies dans de vastes corpus. Ces améliorations permettent de tirer pleinement parti des capacités des LLMs pour la détection d'anomalies textuelles.

7.4 Synthèse

Dans ce chapitre, nous avons mis en place une méthodologie pour l'application des LLMs à la détection d'anomalies textuelles. Nous avons d'abord examiné différentes familles de modèles, telles que GPT, LlaMA, Mistral et Gemini, en analysant leurs architectures de transformeurs, leur taille, et leur processus d'entraînement, pour identifier celles qui conviennent le mieux à notre approche. Les modèles spécifiques seront discutés en détail dans le chapitre suivant consacré aux expériences.

La méthodologie s'est ensuite penchée sur l'utilisation des LLMs en tant que solutionneurs de problèmes, en mettant l'accent sur les composants et la structure des prompts. Nous avons détaillé la manière de concevoir des prompts efficaces, en tenant compte du message du système, des instructions, des exemples, et des stratégies telles que la chaîne de pensée (Chain of Thought). Ces éléments permettent d'orienter les LLMs dans la détection d'anomalies textuelles de manière plus précise et contextuelle.

Enfin, la question de la segmentation des corpus a été abordée comme un point clé pour permettre aux LLMs de traiter efficacement de grandes quantités de données. Grâce à des avancées récentes, telles que l'augmentation de la fenêtre de contexte dans certains modèles comme Gemini, les LLMs peuvent désormais analyser des corpus plus larges et détecter des anomalies avec une plus grande efficacité.

En conclusion, ce chapitre a posé les bases méthodologiques pour l'application des LLMs à la détection d'anomalies textuelles. Le prochain chapitre se consacrera aux expériences concrètes, où nous évaluerons les performances des modèles sélectionnés et leur capacité à résoudre des cas réels de détection d'anomalies textuelles.

CHAPITRE 8

EXPÉRIENCES

Sommaire

8.1 Introduction

Dans les chapitres précédents, nous avons exploré les fondements théoriques et les cadres méthodologiques de la détection d'anomalies textuelles à l'aide des LLMs. Sur cette base, ce chapitre est consacré à un compte rendu détaillé des procédures expérimentales et des analyses menées dans notre recherche. Nos expériences visent à évaluer l'efficacité et l'adaptabilité des LLMs dans le contexte de la détection d'anomalies textuelles.

Ce chapitre est structuré en quatre sections :

1. **Conception et procédure expérimentales** : Cette section décrit la conception générale de nos expériences, en détaillant l'organisation et l'exécution des tests pour évaluer les capacités des LLMs dans la détection d'anomalies.
2. **Modèles et données** : Nous présentons ici les modèles utilisés, leur configuration, ainsi que les corpus sélectionnés pour l'évaluation. La motivation derrière ces choix est également expliquée.

3. **Configuration expérimentale** : Cette section aborde les métriques d'évaluation retenues pour mesurer la performance des LLMs, ainsi que les méthodes de référence pour la comparaison.
4. **Résultats et analyse** : Enfin, nous présentons les résultats de nos expériences, accompagnés d'une discussion approfondie qui analyse et interprète ces résultats dans le but de répondre aux questions de recherche posées au début de cette partie.

Ce chapitre vise non seulement à rapporter les résultats de nos investigations empiriques, mais aussi à réfléchir aux implications de ces découvertes, en discutant à la fois des forces et des limitations de l'utilisation des LLMs dans le contexte de la détection d'anomalies. À travers cette examen détaillé, le lecteur acquerra une appréciation plus profonde des contributions pratiques et théoriques de notre recherche.

8.2 Conception expérimentale

Cette section présente la conception expérimentale de notre étude sur l'efficacité des LLMs dans la détection d'anomalies textuelles. Les expériences sont organisées en trois phases distinctes : les expériences préliminaires, les études d'ablation, et une comparaison systématique avec les méthodes traditionnelles de fouille de données (*Data Mining*, DM). Chaque phase est conçue pour explorer différents aspects de la détection d'anomalies à l'aide des LLMs et pour affiner notre approche en fonction des perspectives acquises.

8.2.1 Expériences préliminaires

La phase préliminaire de nos expériences vise à identifier la configuration optimale pour l'utilisation des LLMs dans la détection d'anomalies, en mettant l'accent sur deux aspects clés : la taille de la fenêtre d'entrée et la composition de la démonstration de la tâche au sein des prompts.

8.2.1.1 Taille de la fenêtre d'entrée

Dans la détection d'anomalies textuelles, la taille de la fenêtre d'entrée, c'est-à-dire le nombre de textes traités par requête, est un facteur crucial en raison de la nature contextuelle de la tâche. Contrairement aux tâches de classification traditionnelles, où les modèles peuvent s'appuyer sur des critères prédéfinis ou des connaissances antérieures pour établir une frontière de décision, la détection d'anomalies repose exclusivement sur l'analyse du jeu de données fourni, sans recours à des informations externes. La décision est donc basée sur une compréhension solide de ce qui est considéré comme « normal » dans les données, ce qui nécessite une fenêtre contextuelle suffisamment large pour capter ces dynamiques.

Dans nos expériences préliminaires, nous avons testé différentes tailles de fenêtres pour évaluer l'impact de la quantité de textes traités par requête sur la capacité du modèle à détecter des anomalies. Les tailles sélectionnées étaient de 10 textes (`w10`), 20 textes (`w20`), 40 textes (`w40`) et 100 textes (`w100`), ajustées en fonction de la capacité contextuelle des modèles utilisés. Ces essais visaient à identifier la configuration optimale qui garantit une détection performante des anomalies, tout en maximisant l'efficacité du traitement en termes de temps et de calcul.

8.2.1.2 Composition de la démonstration

Ces expériences ont pour objectif d'examiner différentes compositions d'exemples de démonstration dans les prompts. Nous avons testé plusieurs configurations, notamment : `1in5` (une anomalie sur cinq exemples), `2in5`, `1in10`, `2in10`, `3in10` et `2in20`. L'analyse s'est concentrée sur trois variables clés :

— **Taille de démonstration** : Déterminer si l'augmentation du nombre total d'exemples améliore systématiquement la performance en termes de précision, rappel ou efficacité globale.
— **Nombre d'exemples positifs** : Examiner si un plus grand nombre d'exemples d'anomalies (positifs) dans le prompt entraîne une amélioration des résultats.
— **Ratio d'anomalies** : Étudier l'impact de l'alignement entre la proportion d'anomalies dans les exemples et la distribution réelle des étiquettes au sein du jeu de données.

8.2.2 Études d'ablation

Les études d'ablation sont conçues pour analyser l'impact des composants individuels d'un système en les supprimant ou en les modifiant de manière sélective. Cette approche aide à isoler les contributions spécifiques de chaque composant à la performance globale du système. Dans notre recherche sur la détection d'anomalies à l'aide des LLMs, l'étude d'ablation a été employée pour disséquer les contributions individuelles de chaque composant dans les prompts, ce qui sert à affiner et optimiser la composition et la structure de ces prompts.

8.2.2.1 Prompt de base

Le prompt de base, tel que décrit dans la section §7.3, consiste en une structure étendue incluant un message du système, des instructions de tâche, une chaîne de pensée (CoT), une démonstration de la tâche (exemples) et une spécification de sortie. Ce prompt complet sert de configuration standard contre laquelle les prompts modifiés sont comparés.

8.2.2.2 Composants évalués

Nos études d'ablation examinent trois principaux composants du prompt (voir la Figure 8.1) :

1. **Message du système et instructions de tâche** : Ces éléments fondamentaux établissent le contexte et définissent la tâche pour le modèle. Nous avons testé des messages du système et des instructions générales, ainsi que des versions spécifiques au scénario, adaptées à des types particuliers d'anomalies et de jeux de données.
2. **Chaîne de pensée** : La CoT est un composant qui modélise explicitement le processus de raisonnement par étapes nécessaire pour aborder la tâche de détection d'anomalies. Nous évaluons l'impact de sa suppression pour déterminer dans quelle mesure le raisonnement par étapes contribue à la performance.
3. **Démonstration de la tâche** :
   — **Standard** : Les exemples mélangés (ou contaminés) contiennent à la fois des échantillons normaux et des anomalies, afin de refléter le scénario typique rencontré dans les tâches de détection d'anomalies.

— **Sans exemples** : Cette variante élimine entièrement la démonstration de la tâche, ce qui permet de tester la capacité intrinsèque du modèle à détecter des anomalies sans exemples explicites.

— **Exemples *Normal-Only*** : Aligné avec le paradigme d'apprentissage semi-supervisé NO, cette configuration utilise uniquement des exemples normaux pour voir si le modèle peut implicitement apprendre à identifier les écarts sans exemples directs d'anomalies.

8.2.2.3 Conditions de l'étude d'ablation

Pour analyser l'impact de ces modifications, nous définissons deux conditions spécifiques pour chaque série d'expériences d'ablation (voir la Figure 8.1) :

— **Condition standard (STD)** : Cela fait référence à la configuration du prompt de base, qui englobe tous les composants mentionnés ci-dessus. Il s'agit du groupe de contrôle dans nos expériences.

— **Condition ablatée (XX-abl)** : Chaque condition d'ablation, désignée par un suffixe indiquant le composant supprimé ou modifié (par exemple, '`CoT-abl (-)`' pour la chaîne de pensée supprimé), explore les effets de l'absence ou de l'altération de ce composant. Cette méthode permet une évaluation précise du rôle de chaque composant dans l'efficacité de la détection d'anomalies par LLMs.

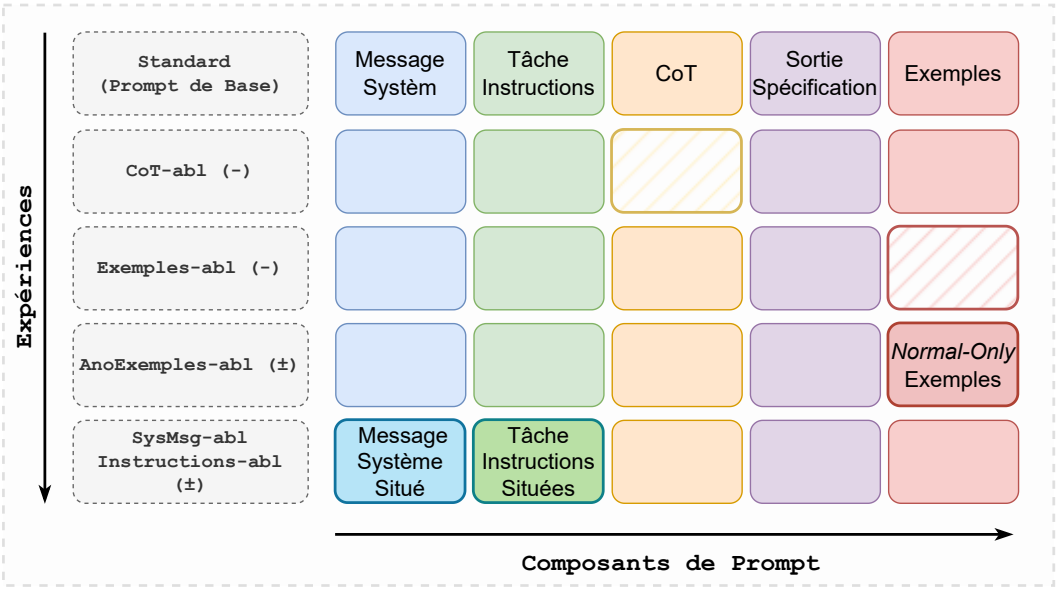


FIGURE 8.1 – Études d'ablation des composants de prompt dans la détection d'anomalies textuelles. Cette figure illustre les différentes configurations de prompt utilisées dans les expériences d'ablation, mettant en évidence les effets de la suppression ou de la modification de composants spécifiques. Les conditions expérimentales comprennent le prompt standard (prompt de base), la suppression de la CoT (CoT-abl), la suppression des exemples (Exemples-abl), l'utilisation d'exemples normaux uniquement (AnoExemples-abl), et la modification du message du système et des instructions (SysMsg-abl et Instructions-abl). Chaque configuration permet d'analyser l'impact de chaque composant sur la performance de détection des anomalies.

8.2.3 Comparaison systématique

La phase finale consiste en une étude comparative entre la détection d'anomalies à l'aide des LLMs et les méthodes traditionnelles de fouille de données (DM) discutées dans la Partie II. Cette étape vise à mettre en lumière les forces et les faiblesses de chaque approche dans des conditions de test identiques en utilisant les mêmes corpus.

À la suite de la comparaison systématique entre les LLMs et les méthodes de DM, nous mènerons une analyse approfondie des erreurs. Cette analyse permettra d'identifier et de catégoriser les erreurs afin de révéler les faiblesses ou limitations spécifiques des LLMs dans la détection d'anomalies textuelles. En examinant les détections ratées et les fausses alertes, nous visons à identifier les conditions sous lesquelles les LLMs sont moins performants et à explorer les améliorations potentielles ou les ajustements nécessaires dans les configurations des modèles et la conception des prompts.

8.3 Modèles et données

8.3.1 Modèles

Parmi les quatre familles de modèles mentionnées dans la section §7.2, à savoir GPT, LlaMA, Mistral et Gemini, nous avons sélectionné cinq modèles pour nos expériences, en nous basant sur leurs performances dans divers classements publics de LLMs. Ces modèles sont : **LlaMA 3-70B**, **Mistral 8×22B**, **GPT-3.5-Turbo**, **GPT-4**, et **Gemini Pro**.

La sélection de ces modèles a été guidée par plusieurs critères essentiels à notre recherche. Parmi eux figuraient : la qualité des modèles, attestée par leurs résultats sur des benchmarks reconnus tels que MMLU [Hendrycks et al., 2021] et MT-bench Zheng et al. [2024] ; la taille de la fenêtre de contexte, qui détermine la longueur maximale du texte que les modèles peuvent traiter en une seule instance ; l'échelle des modèles en termes de nombre de paramètres ; et les conditions de licence sous lesquelles ils sont diffusés [1]

Pour affiner notre sélection, nous avons effectué des tests sur tous les modèles accessibles de chaque famille, en utilisant notamment le corpus TDT2 pour les évaluer. À l'issue de cette phase, ces cinq modèles ont été retenus pour notre évaluation. Lors des expériences préliminaires et des études d'ablation, l'accent a été mis sur LlaMA 3-70B, Mistral 8×22B et GPT-3.5-Turbo, tandis que tous les modèles ont été utilisés dans la phase de comparaison systématique.

Un hyperparamètre important pour ces modèles est la température, fixée constamment à 0,1 pour assurer des sorties stables et prévisibles. Dans les LLMs, la température contrôle le niveau d'aléatoire des réponses ; des valeurs plus basses produisent des sorties plus déterministes et conservatrices, ce qui est essentiel dans la détection d'anomalies pour réduire la variabilité des performances et renforcer la fiabilité des résultats de détection.

1. Voir Annexe B pour la comparaison détaillée de ces critères.

8.3.2 Corpus

Au vu des capacités linguistiques des modèles sélectionnés, notre étude se limite aux corpus en anglais. Nous avons choisi un corpus pour chaque type d'anomalie : **TDT2** pour les anomalies au niveau des thématiques, **Amazon** pour les sentiments déviants, et **OLID** pour les discours de haine et le langage offensant. De plus, pour augmenter le défi, nous avons inclus **TDT2-hard** et **Amazon-hard**, qui présentent des anomalies moins distinguables, comme détaillés dans le Chapitre 4. Pour les expériences préliminaires et les études d'ablation, nous nous sommes principalement concentrés sur les trois corpus standard. Lors de la phase de comparaison systématique, tous les cinq corpus étaient utilisés pour permettre une évaluation complète à travers des jeux de données divers et complexes.

8.4 Configuration Expérimentale

8.4.1 Métriques d'évaluation

Comme indiqué dans la section §7.3, les LLMs sont mieux adaptés à la prise de décision binaire qu'à la génération de scores d'anomalie continus. Par conséquent, les métriques traditionnelles telles que le score AUCROC, appropriées pour évaluer la performance des modèles produisant des sorties continues, ne sont plus applicables ici. Au lieu de cela, notre évaluation repose sur des métriques de classification binaire, incluant la précision, le rappel et le F-score.

Il convient toutefois de noter que ces mesures reflètent les performances d'une méthode à un seuil spécifique, déterminé dans nos expériences par un taux de contamination uniformément fixé. Bien que cette approche offre une base de comparaison cohérente, elle ne reflète pas nécessairement le potentiel global des méthodes de DM, notamment pour des modèles comme XGBOD, dont les performances peuvent varier en fonction des niveaux de seuil.

8.4.2 Méthodes de référence

Pour évaluer la performance des LLMs dans la détection d'anomalies textuelles, nous les avons comparés à des méthodes de DM discutées dans la Partie II , qui servent de références :

1. **Détection d'anomalies non supervisée :** Pour les configurations de LLMs imitant l'apprentissage non supervisé (c'est-à-dire des prompts sans exemples), nous avons comparé leurs performances à celles de méthodes de DM non supervisées telles que ADOD, ALAD, AutoEncoder, COPOD, DIF, HBOS, IForest, KNN, et LOF.
2. **Détection d'anomalies semi-supervisée :** Pour les configurations où les prompts incluent un mélange d'exemples, qui ressemblent aux approches d'apprentissage semi-supervisé et faiblement supervisé, les LLMs ont été évalués en comparaison avec des méthodes de DM comme DeepSAD, DevNet, PreNet, et XGBOD.

Les paramétrages et les réglages des seuils pour ces méthodes de DM de référence ont été rigoureusement maintenus conformément aux directives établies dans la Par-

tie II, assurant une comparaison équilibrée et juste entre différents modèles et méthodes.

8.5 Résultats et analyses

8.5.1 Expériences préliminaires

Nous commençons par les expériences préliminaires visant à optimiser l'utilisation des LLMs dans la détection d'anomalies textuelles, pour évaluer l'impact des différentes tailles de fenêtre d'entrée et des compositions de démonstration sur l'efficacité de la détection d'anomalies. À travers des tests systématiques sur plusieurs modèles et jeux de données, nous cherchons à établir des directives sur le nombre optimal de textes par requête ainsi que sur l'équilibre des types d'exemples dans les prompts pour les expériences suivantes.

8.5.1.1 Taille de la fenêtre d'entrée

Dans nos expériences, nous avons évalué la performance des LLMs à travers différents modèles et ensembles de données, avec des tailles de fenêtre allant de 10 à 100 textes par segment, afin d'identifier la taille de fenêtre optimale pour la détection d'anomalies. Les résultats de cette analyse sont illustrés dans la Figure 8.2.

Tendances générales à travers les modèles et les corpus

- **Score F1** : Le score F1 s'améliore généralement à mesure que la taille de la fenêtre passe de `w10` à `w40` à travers la plupart des jeux de données et modèles. Cela indique qu'une fenêtre de contexte plus large aide les modèles à mieux capturer et comprendre les anomalies, équilibrant ainsi efficacement la précision et le rappel. Cependant, l'amélioration des scores F1 tend à se stabiliser, voire à diminuer légèrement au-delà d'une taille de fenêtre de 40, ce qui suggère qu'un excès de contexte pourrait introduire du bruit ou des informations non pertinentes, perturbant ainsi le modèle plutôt que de l'aider. Par ailleurs, les problèmes opérationnels signalés pour des fenêtres de taille plus importante (supérieures à 50) suggèrent que la performance pourrait se dégrader significativement en raison des difficultés de gestion du contexte.
- **Rappel** : Le rappel diminue progressivement à mesure que la taille de la fenêtre augmente de `w10` à `w40`. Cette baisse pourrait s'expliquer par la dilution du signal anormal lorsque de plus grandes quantités de texte normal sont incluses dans des fenêtres plus larges, rendant ainsi l'identification des anomalies plus subtiles plus difficile pour les modèles.
- **Précision** : La précision augmente initialement avec la taille de la fenêtre, mais commence à diminuer ou à se stabiliser au-delà de `w40`, ce qui indique un seuil à partir duquel l'inclusion de textes supplémentaires peut introduire du bruit ou des informations non pertinentes, réduisant ainsi la précision de la détection d'anomalies.

Taille de fenêtre w40 À la lumière de ces résultats, nous avons décidé de standardiser la taille de la fenêtre à 40 pour tous les modèles et prompts dans les expériences ultérieures, une décision justifiée par plusieurs observations clés :

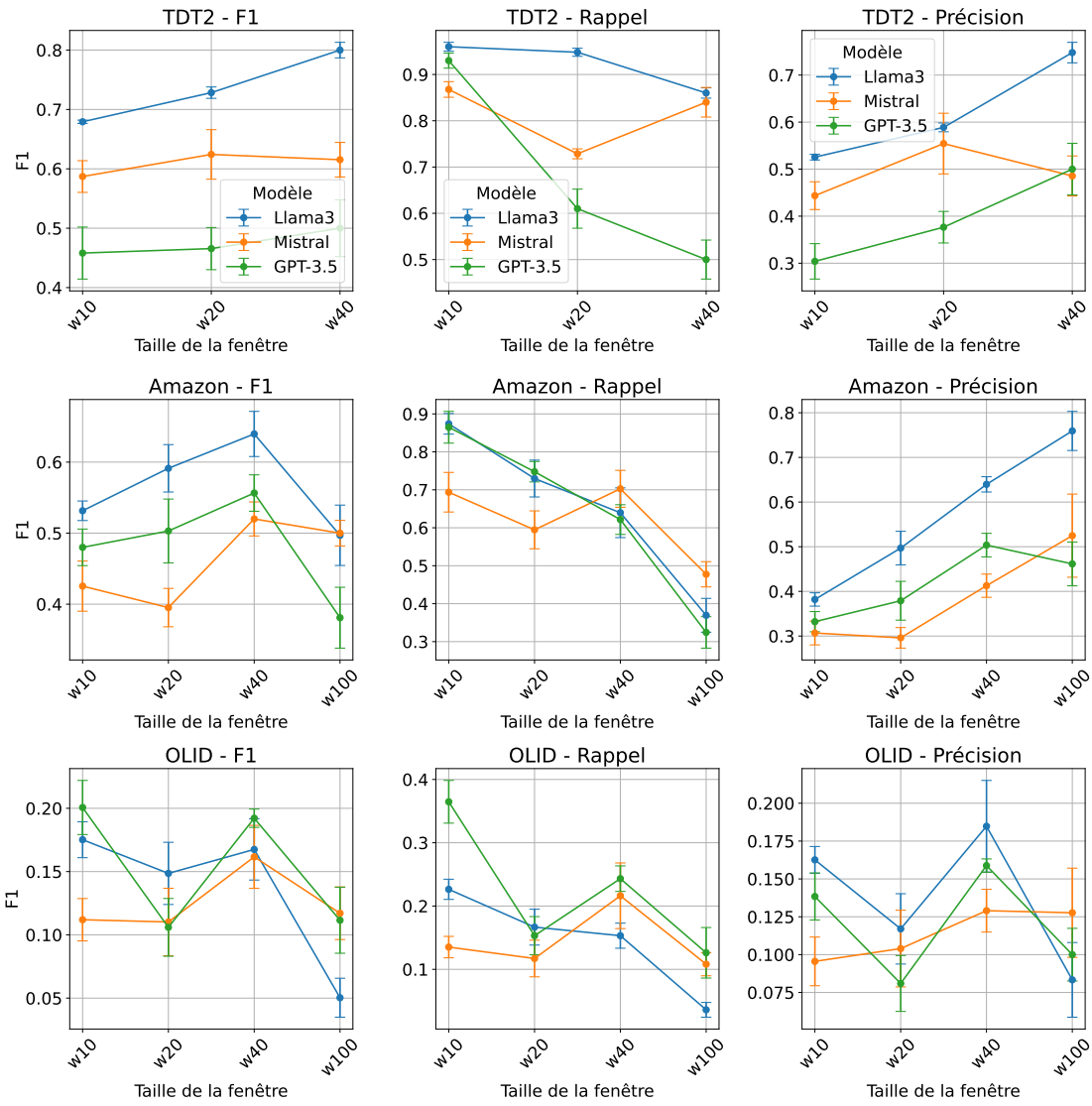


FIGURE 8.2 – Impact de la taille de fenêtre sur la performance de détection d'anomalies textuelles. Cette figure illustre l'effet des différentes tailles de fenêtre (w10, w20, w40) sur les performances (précision, rappel, F1) des modèles (LlaMA 3, Mistral et GPT-3.5) dans la détection d'anomalies textuelles. L'évaluation porte sur trois corpus : TDT2, Amazon, et OLID.

— **Équilibre optimal** : Une fenêtre de 40 textes a constamment offert le meilleur compromis entre l'extension de la compréhension contextuelle par les modèles et le maintien d'une taille d'entrée adaptée à leurs capacités de traitement.

— **Minimisation des erreurs** : Lorsque la taille de la fenêtre dépasse 40, en particulier pour les modèles dotés d'une grande capacité de contexte comme Mistral et Gemini, on observe une augmentation remarquable du nombre d'erreurs opérationnelles. Il s'agit notamment des problèmes de structure JSON mal formée, qui peuvent compromettre l'intégrité des réponses du modèle. De plus, les modèles éprouvent des difficultés à maintenir l'intégralité et l'exactitude de la sortie, comme l'omission de textes et l'absence d'explications pour les anomalies détectées.

— **Dégradation du respect des instructions** : Pour les textes complexes du corpus TDT2, une dégradation rapide de la performance et du respect des instructions a été observée à mesure que la taille de la fenêtre approchait de 80. Cela inclut des comportements inattendus, tels que le non-respect du format de réponse, une généralisation excessive de la norme ou l'identification erronée de la majorité des textes comme anomalies, suggérant une « surcharge cognitive ».

8.5.1.2 Composition de la démonstration

Cette analyse synthétise les résultats obtenus à travers différents modèles et corpus afin d'identifier les tendances générales et les configurations optimales de démonstration (exemples) qui améliorent les capacités des LLMs en matière de détection d'anomalies textuelles (voir la Figure 8.3).

Exemples positifs

— **Amélioration du rappel** : Augmenter le nombre d'exemples d'anomalies dans un total fixe améliore systématiquement le rappel à travers tous les corpus et modèles (par exemple, en passant de `1in5` à `2in5`, et de `1in10` à `3in10`). Cette tendance indique qu'une plus grande exposition aux anomalies, sans submerger le modèle, renforce sa sensibilité et lui permet d'apprendre les caractéristiques des anomalies plus efficacement.

— **Équilibre de la précision** : Alors que l'augmentation des exemples d'anomalies améliore généralement le rappel, elle peut affecter négativement la précision, notamment dans les corpus tels qu TDT2 et Amazon. Dégradation L'équilibre optimal est souvent trouvé dans des configurations comme `2in10`, où il y a suffisamment d'exposition aux anomalies pour améliorer la sensibilité, tout en maintenant un nombre suffisant d'exemples normaux pour contrôler les faux positifs.

Nombre total d'exemples

— **Amélioration de la précision** : Augmenter le nombre total d'exemples améliore généralement la précision, comme on le voit en passant de `1in5` à `1in10`, et de `2in5` à `2in10`. Plus d'exemples fournissent au modèle davantage d'informations contextuelles, lui permettant de différencier plus précisément le contenu anormal du contenu normal. La configuration `2in10` se distingue en améliorant la précision à travers plusieurs corpus, suggérant que ce nombre d'exemples aide le modèle à calibrer efficacement ses frontières de décision dans la détection d'anomalies textuelles.

— **Point de saturation** : Il existe apparemment un point de saturation au-delà duquel des exemples supplémentaires n'améliorent pas substantiellement la performance et peuvent même conduire à des rendements marginaux, comme observé en passant de 10 à 20 exemples dans les configurations comme `2in20`. Cela suggère qu'il existe un nombre optimal d'exemples où les avantages apportés par un contexte supplémentaire s'équilibrent avec les coûts de traitement d'un volume d'informations accru.

Distribution des étiquettes et ratio d'anomalie

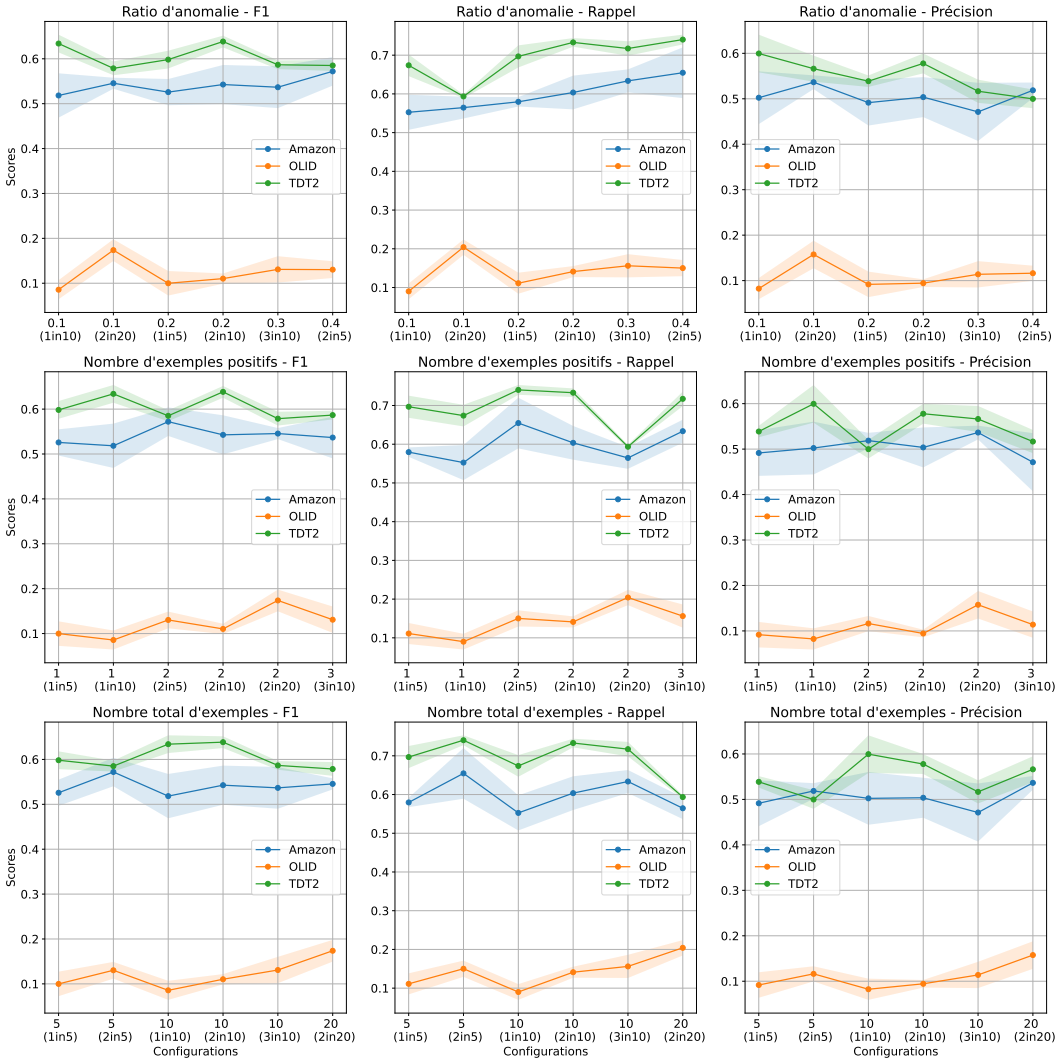


FIGURE 8.3 – Performance moyenne des modèles sur trois corpus (Amazon, TDT2, OLID). Cette figure présente la performance de différents modèles à travers trois corpus, en fonction des métriques (précision, rappel et score F1) et des variables de composition des démonstrations (ratio d'anomalie, nombre d'exemples positifs et nombre total d'exemples). Les variables sont disposées verticalement, tandis que les métriques sont présentées horizontalement, avec l'axe des abscisses organisé en fonction des variables.

— **Distribution réelle** : Contrairement à certains travaux antérieurs comme Min et al. [2022b], nos résultats indiquent que le respect de la distribution réelle des anomalies (par exemple, les configurations `1sur10` et `2sur20`) ne produit pas systématiquement de meilleurs résultats. En fait, ces configurations sont souvent moins performantes que celles avec des ratios d'anomalies ajustés, ce qui indique que les distributions d'étiquettes réelles ne correspondent pas toujours à la composition de démonstration la plus efficace pour la détection d'anomalies.

— **Distributions ajustées** : De légers ajustements du ratio d'anomalies, tels que l'augmentation de la proportion d'exemples d'anomalies (par exemple, `2sur5`), produisent souvent de meilleurs résultats. Cela pourrait être dû à une amélio-

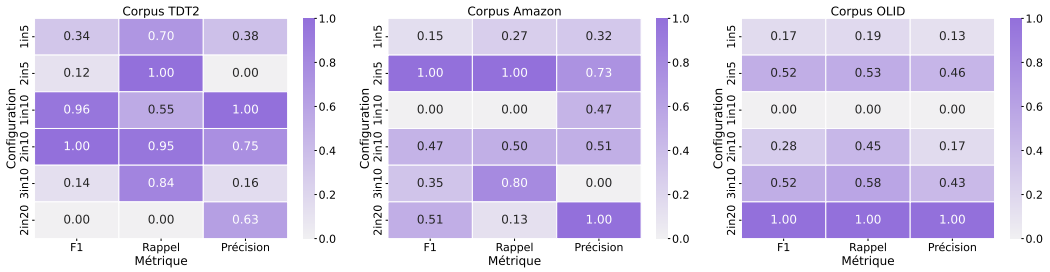


FIGURE 8.4 – Carte thermique des métriques normalisées pour les corpus concernés. Cette carte thermique illustre les métriques normalisées pour chaque corpus, facilitant ainsi la comparaison des performances des différentes configurations sur ces corpus. La normalisation, qui ramène chaque métrique à une échelle de 0 à 1, permet une analyse comparative claire entre les configurations et les métriques. Cette visualisation met en lumière l'efficacité variable des configurations selon les corpus, offrant des pistes stratégiques pour adapter les prompts de détection d'anomalies en fonction des types spécifiques de texte et d'anomalies.

ration de la représentativité des exemples d'anomalies, offrant ainsi au modèle des indices plus clairs pour mieux détecter des anomalies dans des contextes variés, surtout lorsque les anomalies sont subtiles.

Configurations spécifiques aux corpus Pour déterminer la composition de démonstration optimale pour nos expériences, nous avons utilisé une carte thermique de la performance normalisée à travers différents modèles et compositions (Figure 8.4). Cette analyse souligne que les configurations optimales varient considérablement entre les corpus en raison des différences dans les caractéristiques des textes et les types d'anomalies :

— **Corpus TDT2** : La configuration `2sur10` offre le meilleur équilibre, permettant un apprentissage efficace à partir des anomalies tout en maintenant une vue complète du contenu normal, assurant ainsi des capacités de détection robustes.

— **Corpus Amazon** : La configuration `2sur5` excelle à équilibrer le rappel et la précision, la rendant très efficace pour ce corpus.

— **Corpus OLID** : Un nombre total plus élevé avec une distribution réelle, comme `2sur20`, optimise la performance, probablement en raison du besoin de comprendre le contexte de manière détaillée pour identifier des anomalies nuancées comme les discours de haine.

8.5.2 Études d'ablation

Nous procédons maintenant à l'examen des résultats des études d'ablation, conçues pour évaluer les contributions individuelles des différents composants des prompts. Cette analyse porte sur les variations des messages du système et des instructions de tâche, l'inclusion d'une chaîne de pensée (CoT), ainsi que la présence d'exemples. En supprimant ou modifiant ces éléments, nous cherchons à discerner leur impact précis sur la performance du modèle à travers divers scénarios de détection d'anomalies.

Statistique-t et valeur-p Pour évaluer l'impact de ces modifications, nous avons mené une analyse statistique en comparant les métriques de performance moyennes (F1, rappel, précision) entre les conditions standard et ablatées, couvrant différents modèles et corpus. Des tests t appariés ont été utilisés pour calculer les statistiques-t et les valeurs-p, nous permettant ainsi de déterminer la significativité statistique des différences de performance observées entre ces configurations.

— **Statistique-t** : Indique l'ampleur et la direction de la différence entre les deux conditions.
  ⇒ **Valeur positive** : Suggère que la condition standard a surpassé la condition ablatée.
  ⇒ **Valeur négative** : Suggère que la condition ablatée a surpassé la condition standard.
  ⇒ **Ampleur** : Une valeur absolue plus grande indique une différence plus significative entre les conditions.

— **Valeur-p** : Évalue la probabilité que les différences observées se produisent par hasard.
  ⇒ **Valeur-p faible ($< 0,05$)** : Indique une différence statistiquement significative, suggérant que les résultats ne sont probablement pas dus au hasard.
  ⇒ **Valeur-p élevée ($\geq 0,05$)** : Indique une différence qui n'est pas statistiquement significative et qui pourrait être attribuée à une variation aléatoire.

Outre les tests statistiques, nous avons utilisé des diagrammes en boîte pour comparer visuellement la performance des prompts à travers les corpus et les modèles, avec et sans les composants ablatés.

8.5.2.1 Suppression de chaîne de pensée

	LlaMA 3	Mistral	GPT-3.5
t_stat	2,287536	-1,137899	4,554404
p_val	0,028322	0,262892	0,000061

Résultats par Modèle

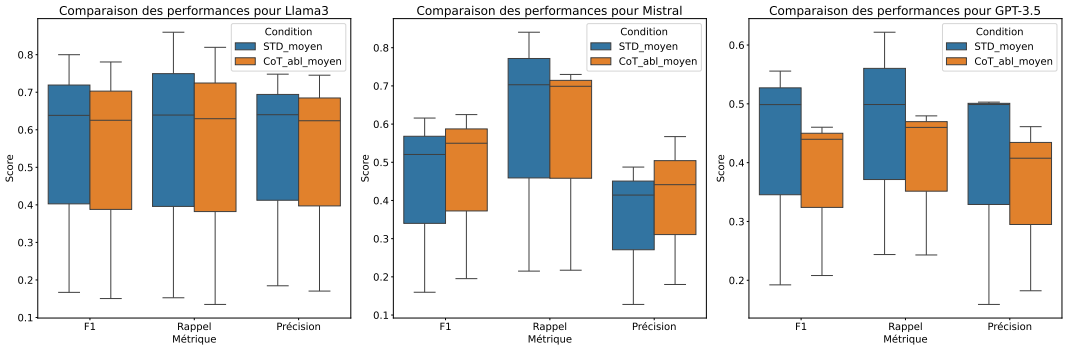
	TDT2	Amazon	OLID
t_stat	2,221692	2,756764	-1,406293
p_val	0,032867	0,009210	0,168454

Résultats par Corpus

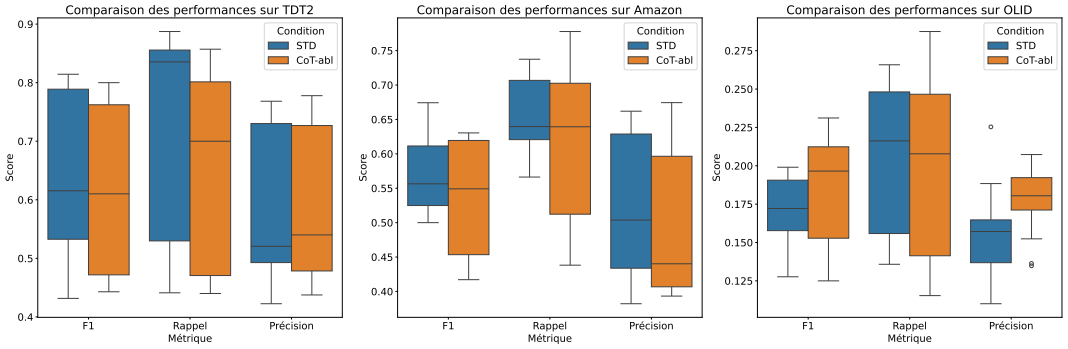
TABLE 8.1 – Analyse statistique des différences de performance dues à l'ablation de CoT à travers les modèles et les corpus. Ces tableaux fournissent un détail statistique, incluant les statistiques-t et les valeurs-p, qui quantifient l'impact de la suppression de CoT des prompts.

Tendances générales

— **Métriques de performance** : Pour la majorité des modèles et des corpus, la présence de CoT dans les prompts améliore généralement la performance en termes de score F1, de rappel et de précision. Les métriques de performance moyennes, illustrées dans la Figure 8.5, montrent des scores plus élevés dans la condition Standard (`STD`) où la CoT est incluse par rapport à la condition sans CoT (`CoT-abl`).



(a) Comparaison de performance à travers les modèles



(b) Comparaison de performance à travers les corpus

FIGURE 8.5 – Comparaison de l'impact de l'inclusion de la chaîne de pensée (CoT) sur les performances à travers différents modèles et corpus. Ces figures illustrent les différences de performance entre les scénarios où la CoT est incluse dans les prompts (en bleu) et ceux où elle est omise (en orange). Chaque figure correspond à un modèle ou à un corpus spécifique, mettant en évidence les gains relatifs en termes de scores F1, de rappel et de précision lorsque la CoT est intégrée.

— **Significativité statistique** : Les statistiques-t et les valeurs-p, rapportées dans le Tableau 8.1, indiquent que les différences de performance sont statistiquement significatives dans la plupart des cas. Les valeurs-p faibles ($< 0,05$) sur plusieurs coprus et modèles suggèrent que les améliorations avec la CoT sont cohérentes et fiables, et ne sont pas dues à des variations aléatoires.

Observations spécifiques aux modèles

— **GPT-3.5** et **LlaMA 3** montrent une baisse évidente des performances à travers différents corpus lorsque la CoT est ablatée, démontrant une dépendance prononcée à la CoT pour des performances optimales.

— En revanche, l'inclusion de la CoT entraîne une baisse de performance pour **Mistral** sur le corpus Amazon, tandis qu'aucune différence importante n'est observée sur les corpus TDT2 et OLID. Globalement, cette baisse de performance de Mistral avec la CoT n'est pas statistiquement significative, comme le montre une valeur-p de 0,263.

Les résultats de notre étude d'ablation démontrent le rôle significatif que joue la CoT dans des performances des LLMs pour la détection d'anomalies textuelles. En favorisant une approche plus structurée à la résolution de problèmes, la CoT per-

met aux modèles de fonctionner avec une plus grande exactitude et fiabilité. Cette observation est solide, confirmée par une significativité statistique constante et des améliorations de performance visibles à travers différents scénarios de test.

8.5.2.2 Suppression de démonstration

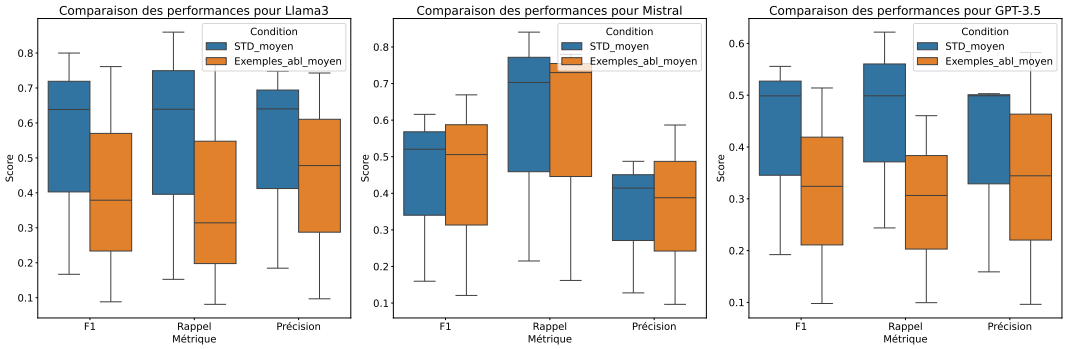
	LlaMA 3	Mistral	GPT-3.5
t_stat	6,874903	0,517934	5,038969
p_val	0,000000	0,607763	0,000014

Résultats par Modèle

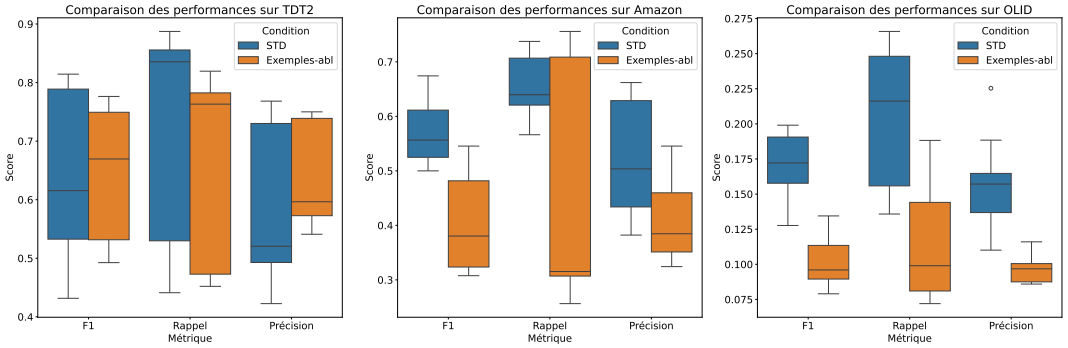
	TDT2	Amazon	OLID
t_stat	-0,262224	7,407911	11,745098
p_val	0,794686	0,000000	0,000000

Résultats par Corpus

TABLE 8.2 – Analyse statistique des différences de performance dues à l'inclusion des exemples dans le prompt à travers les modèles et les corpus.



(a) Comparaison de performance à travers les modèles



(b) Comparaison de performance à travers les corpus

FIGURE 8.6 – Comparaison de l'impact de l'inclusion d'exemples sur les performances à travers différents modèles et corpus. Ces figures illustrent les métriques de performance comparatives à travers différents modèles et corpus, démontrant comment l'inclusion d'exemples améliore la capacité des modèles à détecter efficacement des anomalies.

Tendances générales

— **Métriques de performance** : Comme le montre la Figure 8.6, pour tous les corpus et la plupart des modèles, la présence d'exemples dans les prompts (`STD`) conduit généralement à de meilleures performances par rapport à leur absence (`Exemples-abl`).

— **Significativité statistique** : Les résultats statistiques, résumés dans le Tableau 8.2, soulignent que les différences de performance sont statistiquement significatives dans la plupart des cas, soutenant la conclusion que les exemples sont un élément crucial pour renforcer les capacités de détection des LLMs.

Observations spécifiques aux corpus et modèles

— **GPT-3.5** et **LlaMA 3** présentent une baisse constante de performance pour toutes les métriques lorsque les exemples sont retirés, soulignant leur dépendance aux exemples pour optimiser la performance de détection d'anomalies.

— **Mistral** montre également de légères améliorations avec l'intégration d'exemples. Cependant, une valeur-p de 0,6 indique que la différence n'est pas statistiquement significative, suggérant que cela pourrait être dû au hasard.

— Les corpus **Amazon** et **OLID** montrent des baisses de performance significatives lorsque les exemples sont retirés. Cela indique que les exemples jouent un rôle vital en aidant les modèles à contextualiser et comprendre la tâche de détection d'anomalies dans ces scénarios relativement plus complexes et difficiles.

L'analyse démontre de manière concluante que les exemples dans les pormpts contribuent significativement à l'efficacité des LLMs dans la détection d'anomalies textuelles. En fournissant aux modèles un contexte et des points de référence, les exemples aident à mieux comprendre les nuances et les variations dans les données textuelles, ce qui est crucial pour identifier correctement des anomalies.

8.5.2.3 Altération de démonstration

	LlaMA 3	Mistral	GPT-3.5
t_stat	6,609954	4,077021	3,674156
p_val	0,000000	0,000250	0,000792

Résultats par Modèle

	TDT2	Amazon	OLID
t_stat	3,045666	11,872643	2,361165
p_val	0,004393	0,000000	0,023915

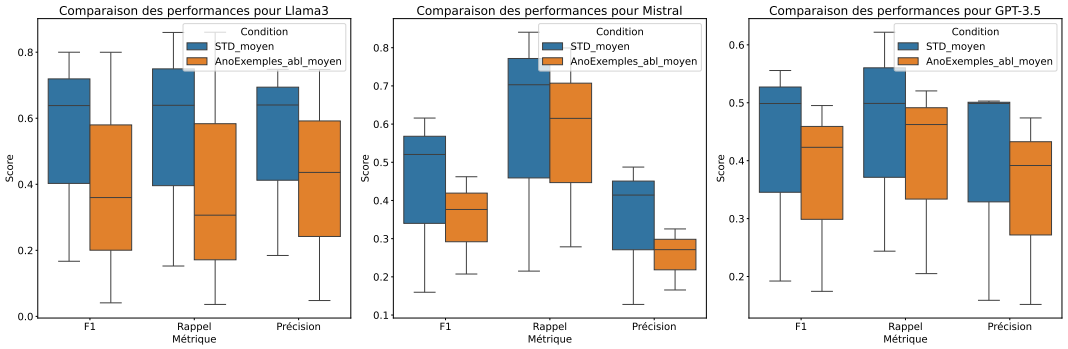
Résultats par Corpus

TABLE 8.3 – Analyse statistique des différences de performance dues à l'inclusion d'exemples d'anomalie (positifs) dans la démonstration à travers les modèles et les corpus. Ces tableaux fournissent une analyse statistique détaillée, incluant les statistiques-t et les valeurs-p, pour quantifier la significativité des différences de performance avec et sans exemples positifs dans la démonstration.

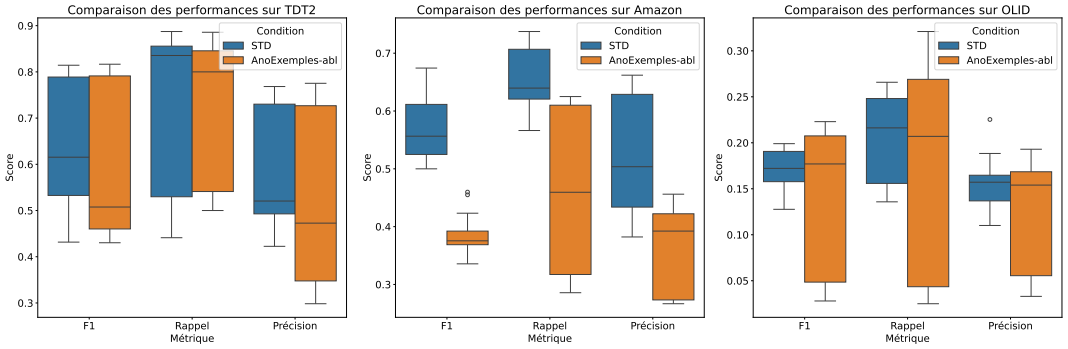
Tendances générales

— **Métriques de performance** : À travers les corpus et les modèles, la condition standard avec des exemples contaminés montre généralement de meilleures performances par rapport à la condition où seulement des exemples normaux sont utilisés (`AnoExemples-abl`). Cela suggère que l'inclusion d'exemples d'anomalies aux côtés des exemples normaux fournit un contexte crucial qui améliore la capacité des modèles à détecter efficacement des anomalies.

— **Significativité statistique** : Les statistiques-t et les valeurs-p, présentées dans le Tableau 8.3, indiquent que les différences de performance sont statistiquement significatives dans la plupart des cas. Des valeurs-p faibles ($< 0,05$) à travers de multiples corpus et modèles suggèrent que les améliorations avec



(a) Comparaison de performance à travers les modèles



(b) Comparaison de performance à travers les corpus

FIGURE 8.7 – Comparaison de l'impact de l'inclusion d'exemples d'anomalies (positifs) sur les performances à travers différents modèles et corpus. Ces figures montrent comment l'utilisation d'exemples contaminés, comparée à celle d'exemples uniquement normaux, affecte les capacités de détection des différents modèles sur divers corpus, en mettant en évidence les différences dans les scores F1, le rappel et la précision.

l'inclusion d'exemples contaminés sont cohérentes et fiables, et ne sont pas dues à une variation aléatoire.

L'inclusion d'exemples normaux et d'anomalies dans les prompt améliore significativement la performance des LLMs dans les tâches de détection d'anomalies à travers divers jeux de données. Cela suggère que les exemples contaminés fournissent un contexte essentiel qui aide les modèles à différencier plus efficacement entre le contenu normal et anormal.

8.5.2.4 Messages du système et instructions spécifiques au scénario

Tendances générales

— **Métriques de performance** : La Figure 8.8 montre des réactions variées des modèles à la suppression des éléments spécifiques au scénario dans les messages du système et les instructions de tâche. Alors que certains modèles enregistrent une baisse de performance, d'autres affichent peu de changements.

— **Significativité statistique** : Les résultats des tests statistiques, tels que reflétés dans le Tableau 8.4, indiquent différents niveaux de significativité : GPT-3.5 montre une baisse de performance statistiquement significative lorsque les élé-

	GPT-3.5	Mistral	LlaMA 3
t_stat	2,695800	2,019531	1,361118
p_val	0,020812	0,068471	0,200708

TABLE 8.4 – Analyse statistique de l'impact des instructions spécifiques au scénario à travers les modèles. Ce tableau détaille les statistiques-t et les valeurs-p pour chaque modèle, quantifiant la significativité des différences observées dues à la présence ou à l'absence d'éléments spécifiques au scénario.

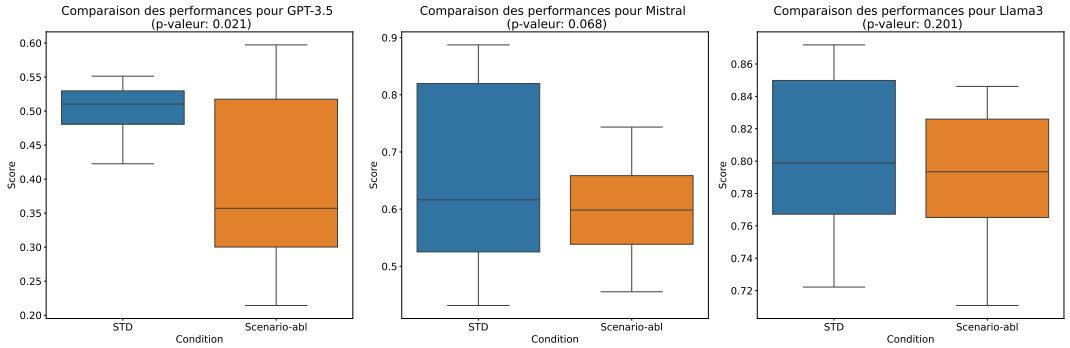


FIGURE 8.8 – Effet des instructions spécifiques au scénario sur comparaison des performances de LLMs dans le corpus TDT2. Cette figure illustre la comparaison des performances de différents LLMs avec et sans éléments spécifiques au scénario dans le prompt, soulignant la variabilité de leur dépendance à ces éléments pour une performance optimale.

ments spécifiques au scénario sont retirés (valeur-p : 0.021), suggérant que ces éléments sont cruciaux pour une performance optimale ; Mistral présente une valeur-p marginale (0.068), indiquant une tendance presque significative où la spécificité du scénario pourrait modérément influencer la performance ; LlaMA 3 affiche une valeur-p élevée (0.201), suggérant qu'il n'y a pas d'impact significatif de l'inclusion d'éléments spécifiques au scénario.

L'inclusion d'éléments spécifiques au scénario dans les messages du système et les instructions de tâche améliore la performance pour certains modèles mais n'est pas universellement bénéfique. Cela suggère que, bien que les éléments spécifiques au scénario puissent aider à contextualiser les tâches, leur utilité est dépendante du modèle et influencée par les mécanismes sous-jacents de chaque modèle.

8.5.3 Analyse comparative

8.5.3.1 LLMs contre méthodes de DM non supervisées

Performance globale Les résultats de nos expériences (voir la Figure 8.9) révèlent que, globalement, les LLMs surpassent les méthodes traditionnelles de DM sur les trois corpus standards, notamment TDT2, qui se concentre sur les anomalies au niveau des thématiques. Cette supériorité est principalement due à des scores de précision plus élevés des LLMs. Cependant, la performance des LLMs décline significativement sur les anomalies textuelles plus difficiles (sentiments déviants et discours haineux), où ils sont généralement moins performants par rapport aux méthodes de DM.

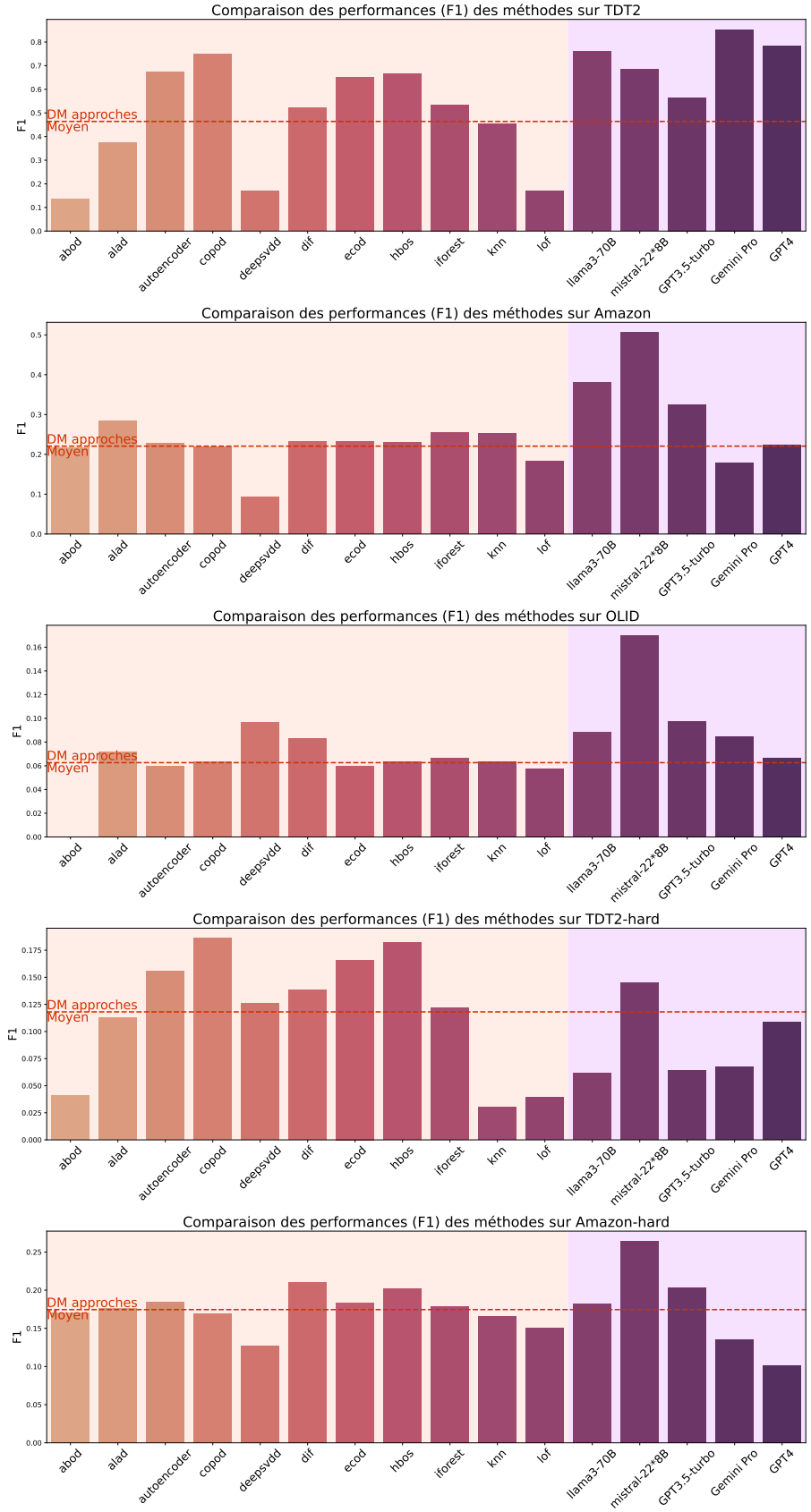


FIGURE 8.9 – LLMs à l'aide des prompts sans démonstration vs méthodes traditionnelles de fouille de données non supervisées. Cette figure compare les scores F1 des LLMs avec prompts sans démonstration (en violet clair) et des méthodes traditionnelles de fouille de données non supervisées (DM, en rouge clair) sur cinq corpus : TDT2, Amazon, OLID, TDT2-hard, et Amazon-hard. Le score F1 moyen des méthodes de DM pour chaque corpus est représenté par une ligne rouge.

Observations spécifiques aux modèles

— **Mistral** démontre un avantage constant dans un contexte sans échantillons annotés (scénarios non supervisés). Il excelle sur tous les critères et jeux de données, surpassant non seulement les méthodes de DM mais aussi d'autres LLMs.

— **LlaMA 3** et **GPT-3.5** suivent la tendance générale à exceller dans les corpus standards en termes de précision et de scores F1, mais échouent sur les corpus avec des paires anomalie-normalité moins distinguables.

— **Gemini Pro** montre une compétence unique dans la détection d'anomalies au niveau des thématiques et des événements au sein des corpus TDT2 et TDT2-hard, indiquant sa capacité spécialisée en analyse thématique. Cependant, sa performance déclinent dans d'autres corpus, mettant en évidence une éventuelle étroitesse d'application au-delà des types d'anomalies spécifiques.

— **GPT-4**, malgré sa performance de pointe dans de nombreuses tâches de TALN, ne montre un avantage que dans l'ensemble de données TDT2. Dans d'autres corpus, il est généralement moins performant par rapport aux méthodes de DM traditionnelles, principalement à cause de son faible rappel. Les résultats révèlent une inclination conservatrice évidente de GPT-4 en l'absence d'exemples, qui atteint une haute précision mais manque de nombreuses anomalies.

Observations spécifiques aux corpus

— Pour le corpus **TDT2**, qui traite des anomalies au niveau des thématiques, les LLMs performent à la hauteur des méthodes de DM à base de statistiques-probabilités et surpassent significativement d'autres méthodes de DM. Cette performance équilibrée en précision et rappel souligne le potentiel des LLMs dans la détection d'anomalies liées aux thématiques dans les scénario relativement simple.

— Pour **TDT2-hard** et **Amazon-hard**, deux corpus particulièrement difficiles pour les méthodes de DM, les LLMs se montrent encore plus vulnérables. La nature complexe de ces ensembles de données, où les anomalies sont moins distinguables et plus nuancées, met davantage en évidence les limites des LLMs dans la gestion de tâches de détection d'anomalies complexes et subtiles. Les faibles scores de rappel observés dans ces corpus difficiles indiquent que les LLMs, malgré leurs capacités avancées, manquent souvent une proportion significative d'anomalies.

8.5.3.2 LLMs contre méthodes de DM semi-supervisées et faiblement supervisées

Performance globale

— Dans les corpus plus simples tels que TDT2 et Amazon, les LLMs démontrent une performance supérieure comparée aux méthodes traditionnelles de DM lorsque $\gamma = 10$, et restent compétitifs ou légèrement meilleurs lorsque $\gamma = 40$. Cependant, dans les corpus plus complexes comme OLID et les versions difficiles de TDT2 et Amazon, les LLMs montrent un déclin de performance plus prononcé. Cette tendance met en lumière leurs limitations dans la gestion d'anomalies complexes et subtiles où la distinction entre le texte normal et anormal est moins claire.

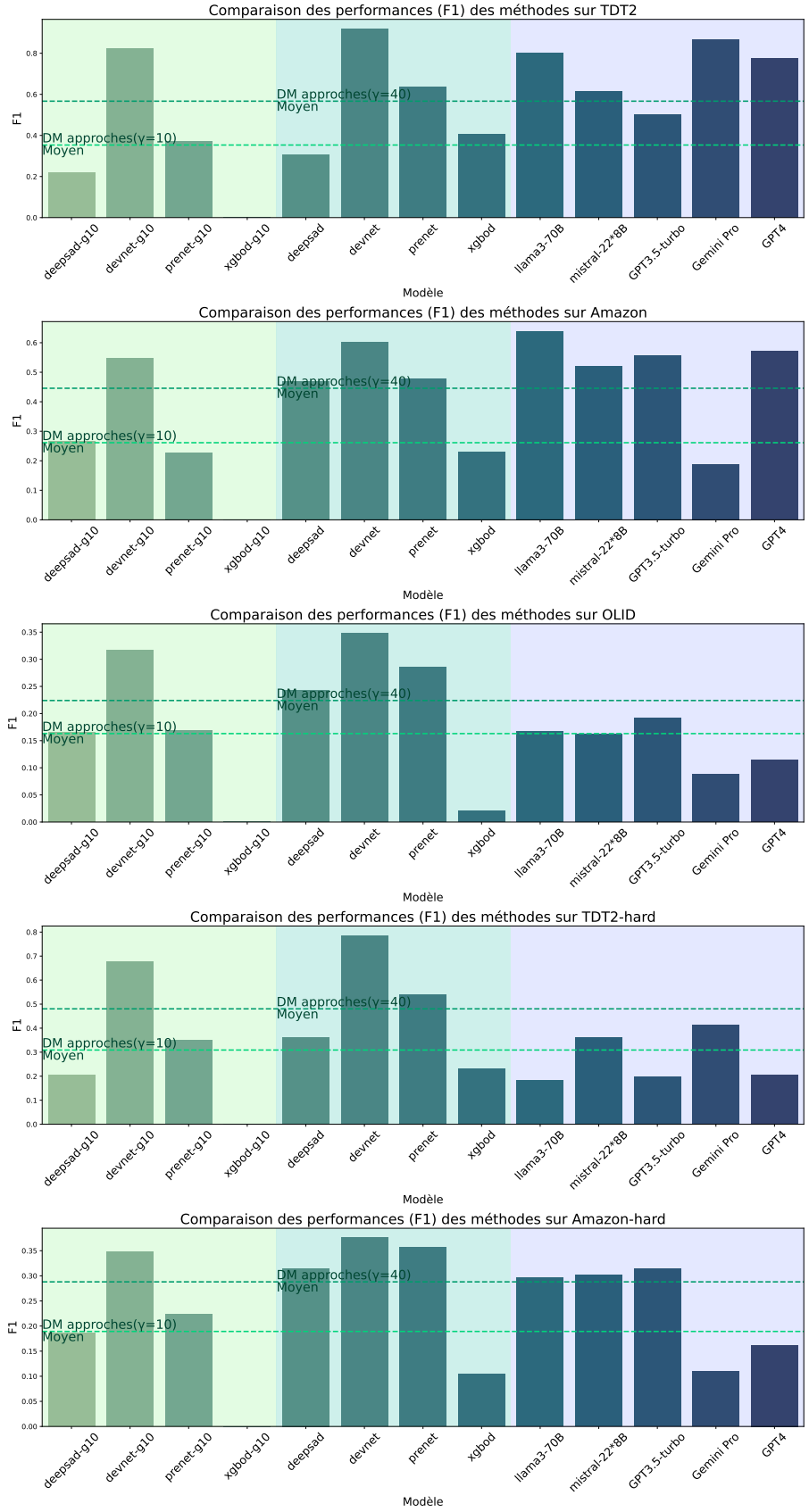


FIGURE 8.10 – LLMs à l'aide des prompts avec démonstration vs méthodes traditionnelles de fouille de données semi-supervisées et faiblement supervisées. Cette figure compare les scores F1 des LLMs à l'aide des prompts démonstratifs (en bleu foncé) et des méthodes traditionnelles de fouille de données semi-supervisées et faiblement supervisées (en vert clair pour $\gamma_l = 10$, en bleu clair pour $\gamma_l = 40$) à travers cinq corpus. Les scores F1 moyens des méthodes de DM pour chaque corpus sont indiqués par les lignes vertes claires pour $\gamma = 10$ et les lignes vertes foncées pour $\gamma = 40$.

— Les LLMs ont manifesté une tendance conservatrice évidente en l'absence d'exemples. En revanche, dans les scénarios avec des prompt de type *few-shot*, les LLMs, particulièrement GPT-3.5, montrent un changement vers un rappel plus élevé mais au détriment de la précision, indiquant une approche de détection d'anomalies plus agressive. Un tel comportement est indicatif de la réponse adaptative des LLMs à l'inclusion d'exemples explicites, qui pourrait modifier leurs seuils de détection.

Observations spécifiques aux modèles

— **LlaMA 3** apparaît souvent parmi les modèles qui offrent les meilleures performances. Ce modèle équilibre le rappel avec un niveau acceptable de précision, le rendant bien adapté aux scénarios où aussi bien les faux positifs que les faux négatifs ont des conséquences significatives.

— **GPT-3.5** et **Mistral** suivent de près Llama3 avec un rappel spécifiquement élevé, ce qui les rend idéaux pour des environnements à forts enjeux où le défaut de détection d'une anomalie pourrait avoir de graves conséquences.

— La performance de **Gemini Pro** et de **GPT-4** varie considérablement. Gemini Pro montre un profil de performance distinct, excellant spécifiquement dans la gestion des anomalies au niveau des thématiques au sein du corpus TDT2 et sa variante plus difficile. Cependant, son efficacité diminue avec d'autres types d'anomalies à travers différents corpus. GPT-4 présente une performance moyenne dans les coprus plus simples mais rencontre des difficultés notables dans les jeux de données complexes. Cette variation de performance souligne les défis potentiels de la capacité du modèle à généraliser ses capacités de détection à des types d'anomalies plus nuancés ou moins distincts.

8.5.4 Analyse des erreurs

Dans l'analyse suivante, nous visons à examiner les erreurs de prédiction rencontrées lors des expériences avec les LLMs et à explorer les origines de ces erreurs. Notre attention se portera sur deux types principaux d'erreurs (Figure 8.11a) :

— **Raté** : Cette erreur se produit lorsque des anomalies sont incorrectement identifiées comme du texte normal.

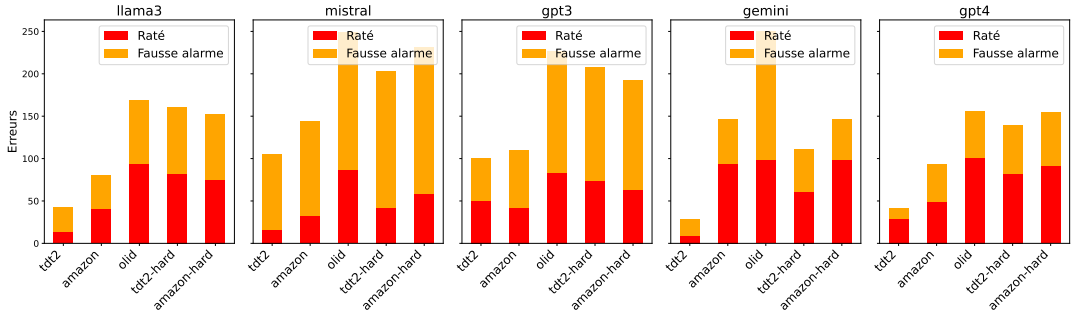
— **Fausse Alarme** : Cette erreur survient lorsque des textes normaux sont à tort déclarés comme des anomalies.

Un avantage des LLMs dans la détection d'anomalies est leur capacité à fournir des sorties qui incluent les normes établies et des explications pour les décisions, améliorant l'interprétabilité et la transparence du processus de prise de décision. En examinant les normes et les explications produites par les modèles, nous pouvons efficacement retracer les racines de ces erreurs (Figure 8.11), améliorant ainsi la performance pour les études futures.

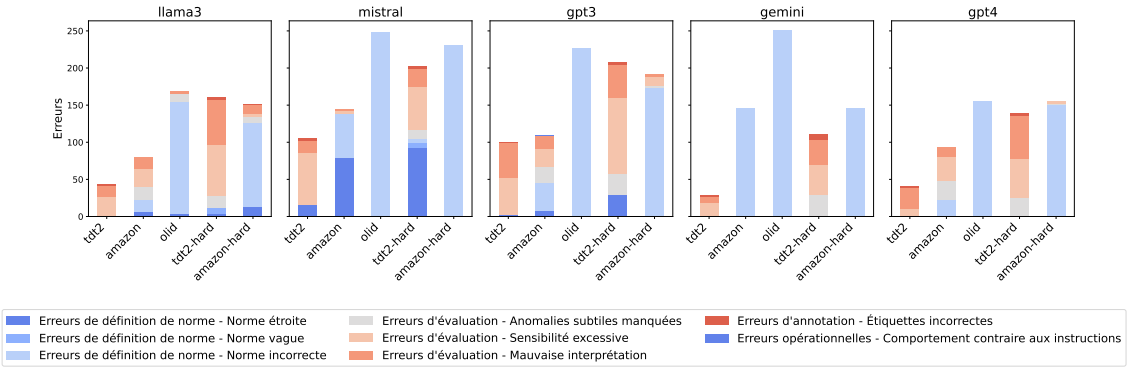
8.5.4.1 Analyse des origines des erreurs

Problèmes d'établissement des normes Ces problèmes surviennent lors de la phase d'établissement des normes et peuvent être classés en trois sous-catégories :

1. **Norme définie de manière extensive** : Les normes sont trop inclusives, englobant des variations qui devraient être considérées comme des anomalies.



(a) Distribution des erreurs spécifiques au modèle à travers le corps



(b) Analyse de l'origine des erreurs à travers les modèles et les corpus

FIGURE 8.11 – Analyse des types et des origines des erreurs dans la détection d'anomalies avec LLMs à travers divers corpus. La sous-figure (a) illustre la distribution des erreurs (raté et fausses alarmes) pour chaque modèle dans différents corpus, ce qui donne un aperçu de la sensibilité et de la spécificité de chaque modèle dans la détection d'anomalies. La sous-figure (b) examine plus en détail les causes des erreurs, en les regroupant en plusieurs catégories : problèmes liés à l'établissement des normes (normes étroites, larges et incorrectes), difficultés d'évaluation (anomalies subtiles non détectées, sensibilité excessive et interprétation erronée) et problèmes opérationnels (étiquettes incorrectes et non-respect des instructions).

Cela entraîne des taux de ratés plus élevés, car ces anomalies sont incorrectement classées comme normales.

❖ **Exemple**

⇒ **Modèle et texte** : LlaMA 3, TDT-2-hard [TEXTE 605]

⇒ **Norme définie** : « *Reports on the Asian financial crisis, including economic reforms, IMF bailouts, and political unrest in countries such as Indonesia, Japan, and South Korea.* »

⇒ **Norme réelle** : « *Economic crisis in Asia. Topic related stories include- IMF bail-out and U.S. involvement, ripple effect in U.S. and other stock markets, impact on international businesses, impact of people's daily life.* »

⇒ **Anomalie réelle** : « *Anti-government protests and violence motivated by the economic crisis.* »

⇒ **Problème** : La définition très vague a entraîné des chevauchements entre les textes normaux et anormaux, ce qui a empêché la détection de certaines anomalies telles que les textes concernant les bouleversements politiques.

2. **Norme définie de manière étroite** : Les normes sont trop spécifiques„ ce qui entraîne des taux de fausses alarmes plus élevés, car les textes normaux qui dévient légèrement sont incorrectement signalés comme des anomalies.

❖ **Exemple**

⇒ **Modèle et texte** : GPT-4, TDT2-hard, [TEXTE 964]

⇒ **Norme définie** : « *Indonesia's economic crisis and its global impact.* »

⇒ **Norme réelle** : « *Economic crisis in Asia.* »

⇒ **problème** : Les textes abordant la crise économique dans des pays non explicitement mentionnés dans la norme étroite (comme le Japon et la Chine) ont été déclarés à tort comme des anomalies.

3. **Norme mal définie** : Les normes ne reflètent pas fidèlement les patterns principaux du corpus, conduisant à des interprétations erronées et à des classifications incorrectes des textes normaux et anormaux.

❖ **Exemple**

⇒ **Modèle et texte** : Gemini Pro, OLID, [TEXTE 209]

⇒ **Norme définie** : « *Tweets discussing politics, opinions, and personal interactions.* »

⇒ **Norme réelle** : « *Comments using neutral or non-offensive language.* »

⇒ **Problème** : La mauvaise définition de la norme a conduit à des erreurs de classification basées sur la nature du contenu.

Problèmes d'évaluation Ces problèmes surviennent lors du processus d'évaluation des déviations par rapport aux normes établies et peuvent conduire à des ratés et des fausses alarmes :

1. **Anomalies nuancées ratées** : Le modèle ne parvient pas à détecter des écarts subtils mais significatifs indiquant des anomalies, ce qui se traduit par des taux ratés élevé, avec des anomalies classées incorrectement comme normales.

❖ **Exemple**

⇒ **Modèle et texte** : LlaMA 3, Amazon-hard, [TEXT 645]

⇒ **Norme définie** : « *Product reviews with positive or neutral sentiment discussing features, fit, and quality.* »

⇒ **Contenu** : « *I love these pants, but they started coming until stitched after a few wears and washes ...* »

⇒ **Problème** : Le modèle n'a pas détecté les implications légèrement négatives, classant la critique comme normale en raison de son incapacité à identifier le glissement subtil du sentiment.

2. **Sensibilité excessive** : Le modèle est trop sensible aux écarts mineurs et signale incorrectement les textes normaux comme des anomalies.

❖ **Exemple**

⇒ **Modèle et texte** : LlaMA 3, TDT2-hard [TEXT 126]

⇒ **Norme définie** : « *Discussions on the economic crisis in Asia.* »

⇒ **Explication** : « *Discussion about Malaysian banks raising lending rates.* »

⇒ **Problème** : Le modèle a marqué ce détail financier spécifique comme une anomalie, reflétant une réponse trop sensible à une variation normale dans le cadre de la discussion économique plus large.

3. **Interprétation et classification erronées** : Le modèle ne comprend pas correctement le contexte ou la sémantique du texte, conduisant à une détection d'anomalie incorrecte.

❖ **Exemple**

⇒ **Modèle et texte** : Mistral, TDT2, [TEXT 152]

⇒ **Norme définie** : « *Texts discussing the economic crisis in Asia, impacts on various countries.* »

⇒ **Contenu actuel** : « *Report on an avalanche near the Italian border.* »

⇒ **Prédiction** : `normal (0)`

⇒ **Problème** : Le modèle a classé incorrectement le rapport comme lié à la crise économique en raison de son incapacité à interpréter correctement le contexte et le sujet du texte.

Problèmes d'annotation et de fonctionnement Outre les problèmes principaux liés aux phases de détection, des erreurs peuvent également survenir en raison de facteurs moins courants, tels que l'annotation incorrecte du corpus et le comportement inattendu du modèle.

1. **Erreurs d'annotation** : Les étiquettes de vérité terrain sont inexactes ou ne correspondent pas au contenu du texte, entraînant des écarts dans l'évaluation.

❖ **Exemple**

⇒ **Modèle et texte** : GPT-4, TDT-2, [TEXT 126]

⇒ **Contenu** : « *A british aircraft carrier and an auxiliary ship are moving through the suez canal en route to the persian gulf ...* »

⇒ **Étiquette** : « *Asian Economic Crisis (Normal)* »

⇒ **Problème** : Le contenu du texte a été incorrectement étiqueté comme lié à la crise économique asiatique, entraînant des erreurs d'évaluation.

2. **Comportement contraire aux instructions de tâche** : Il arrive que le modèle se comporte d'une manière contraire aux instructions de la tâche. Par exemple, bien que la norme soit correctement établie et l'écart par rapport à la norme identifié, le texte est toujours classé comme normal.

❖ **Exemple**

⇒ **Modèle et texte** : GPT-3.5, Amazon, [TEXT 126]

⇒ **Prédiction** : `anormal (1)`

⇒ **Explication** : « *Positive review with sizing up mentioned, not a deviation from the norm.* »

⇒ **Problème** : Malgré l'identification de la critique positive comme correspondant à la norme, le texte a été incorrectement classé comme une anomalie, indiquant un échec dans le processus de décision.

8.5.4.2 Analyse des erreurs spécifiques aux modèles et aux corpus

LlaMA 3 présente une distribution relativement équilibrée entre les ratés et les fausses alarmes dans la plupart des cas, démontrant un niveau optimal de sensibilité et de spécificité. Toutefois, cet équilibre est rompu sur les corpus TDT2 et TDT2-hard, où une augmentation notable des fausses alarmes est observée. Cette hausse est liée à une sensibilité excessive du modèle lors de l'évaluation, probablement due à

une surexposition aux anomalies potentielles via les prompts. Ceci indique un besoin d'ajustement des prompts pour mieux distinguer entre les variations normales et les anomalies véritables, en particulier pour les jeux de données caractérisés par des variations normales complexes.

Mistral est marqué par des taux élevés de fausses alarmes dans différents corpus, principalement à cause de problèmes dans l'établissement des normes. Ce modèle adopte souvent une définition trop étroite des normes, comme le montre sa gestion des corpus Amazon et TDT2-hard. Cette étroitesse provient probablement d'un ajustement excessif aux exemples normaux dans les prompts, amenant Mistral à percevoir toute légère déviation comme anomale. La persistance de ce problème suggère que l'approche de Mistral pour définir les normes est trop restrictive, limitant ainsi sa capacité à se généraliser face aux variations du monde réel et entraînant de fréquentes fausses alarmes.

GPT-3.5 se distingue par une approche de prédiction agressive, menant à des taux plus élevés de fausses alarmes dans tous les corpus évalués. La sensibilité de ce modèle durant la phase d'évaluation, bien que potentiellement utile pour détecter des anomalies subtiles, se manifeste souvent par un excès de vigilance, où même des écarts mineurs sont signalés comme des anomalies potentielles. Ajuster la sensibilité des prompts pourrait aider à équilibrer les capacités de détection de GPT-3.5.

Contrairement aux autres modèles, **GPT-4** adopte une approche plus conservatrice, avec une tendance plus marquée à ne pas détecter les anomalies plutôt qu'à signaler à tort des textes normaux comme anormaux. Cette approche conservatrice est particulièrement évidente dans le corpus TDT2, où elle contraste avec d'autres modèles qui affichent un nombre plus élevé de fausses alarmes.

Les corpus **OLID**, **Amazon**, et **Amazon-Hard** sont particulièrement difficiles en termes d'établissement des normes. Définir ce qui constitue un comportement normal dans des contextes chargés de sentiments déviants et de discours haineux nécessite une compréhension approfondie de l'usage nuancé de la langue. La difficulté à établir des normes efficaces via les prompts actuelles souligne le besoin d'une ingénierie plus sophistiquée des prompts pour capturer les subtilités complexes de la langue inhérentes à ces contextes.

Dans les corpus **TDT2-Hard** et **Amazon-Hard**, où les anomalies sont moins distinguables, le défi de rater des anomalies subtiles est particulièrement prononcé. Les prompts actuels pourraient ne pas capturer suffisamment les nuances nécessaires pour détecter ces écarts discrets, ce qui suggère un axe d'amélioration possible dans la conception des prompts pour accroître la sensibilité aux détails plus fins sans pour autant augmenter le bruit des fausses alarmes.

8.6 Synthèse

Ce chapitre a détaillé les expériences menées pour évaluer l'efficacité des LLMs dans la détection d'anomalies textuelles. Les expériences se sont articulées autour de trois axes : les tests préliminaires, les études d'ablation, et une comparaison systématique avec des méthodes traditionnelles de fouille de données.

Dans les **tests préliminaires**, nous avons exploré différentes configurations de taille de fenêtre et de composition de prompts pour optimiser l'utilisation des LLMs.

Il en ressort que les fenêtres contextuelles plus larges (jusqu'à 40 textes) et des démonstrations riches en exemples d'anomalies améliorent la performance des modèles, tant en termes de précision que de rappel.

Les **études d'ablation** ont mis en lumière l'importance de certains composants des prompts, tels que la chaîne de pensée (CoT) et les exemples. L'inclusion de ces éléments a permis d'augmenter significativement la capacité des LLMs à détecter des anomalies textuelles complexes.

Enfin, la **comparaison systématique** avec les méthodes de fouille de données a montré que les LLMs surpassent largement les méthodes traditionnelles dans des scénarios où les données étiquetées sont totalement absentes. Grâce à des descriptions simples de la tâche ou à des instructions étape par étape, les LLMs parviennent à résoudre des problèmes de manière plus efficace que les méthodes de fouille de données classiques. De plus, dans des scénarios où nous disposons de très peu de données étiquetées, les LLMs démontrent une capacité notable à bien généraliser à partir d'un petit nombre d'exemples. Il convient toutefois de noter que, face à des scénarios plus complexes où les anomalies sont plus subtiles, les performances des deux types de méthodes, qu'il s'agisse des méthodes de fouille de données ou des LLMs, tendent à se dégrader. Néanmoins, cette dégradation est plus marquée pour les LLMs, ce qui souligne l'importance d'adapter les prompts et stratégies de détection pour ces contextes plus difficiles.

Conclusion de la troisième partie

Dans cette partie, nous avons cherché à adapter et à appliquer les LLMs pour répondre aux exigences spécifiques de la détection d'anomalies dans les données textuelles. Cette exploration nous a permis de mieux comprendre comment les LLMs, avec leurs capacités émergentes, peuvent être intégrés dans des pipelines de détection d'anomalies textuelles, offrant des solutions plus avancées et performantes que les méthodes traditionnelles.

* * *

Le Chapitre 6 a exposé les fondements théoriques des LLMs, en présentant leur architecture, leur échelle, et leurs capacités émergentes. Ce chapitre a permis de comprendre comment ces modèles peuvent être pré-entraînés sur de grandes quantités de données non structurées pour ensuite être adaptés à diverses tâches, y compris la détection d'anomalies textuelles. Nous avons mis en avant les avantages des modèles autorégressifs et discuté du rôle central des prompts dans la direction et la performance des LLMs.

Le Chapitre 7 s'est centré sur la méthodologie développée pour exploiter les LLMs dans la détection d'anomalies textuelles. Ce chapitre a mis l'accent sur l'utilisation des LLMs comme solutionneurs de problèmes, en particulier via la conception de prompts. Nous avons exploré comment structurer les prompts pour orienter les modèles dans des scénarios complexes, en tenant compte de l'importance de la segmentation des données et du contexte des corpus textuels. Ce travail méthodologique a démontré que l'optimisation des prompts et la gestion du contexte textuel sont des facteurs critiques pour maximiser la performance des LLMs dans la détection d'anomalies.

Le Chapitre 8 a rapporté les résultats des expériences menées avec les LLMs, en comparant leur performance à celle des méthodes traditionnelles de fouille de données. Les résultats ont montré que les LLMs sont très performants dans des tâches simples de détection d'anomalies, mais qu'ils rencontrent certaines limites face à des anomalies plus subtiles. L'étude des performances a révélé que certains composants, comme la chaîne de pensée et les exemples, sont cruciaux pour maximiser l'efficacité des LLMs dans des contextes plus complexes.

* * *

Les LLMs peuvent-ils exceller dans la détection d'anomalies textuelles et comment se comparent-ils aux méthodes traditionnelles ? Les résultats expérimentaux ont montré que les LLMs surpassent souvent les méthodes de fouille de données traditionnelles dans la détection d'anomalies textuelles, en particulier pour les anomalies bien définies et dans les contextes où les données annotées sont inexistantes. Leur capacité à généraliser sans supervision directe les rend particu-

lièrement efficaces dans ces scénarios. Cependant, face à des jeux de données plus complexes ou à des anomalies plus subtiles, leurs performances se dégradent davantage par rapport aux méthodes de fouille de données, mettant en évidence la nécessité d'optimisations supplémentaires pour les adapter à des tâches plus nuancées.

Quel modèle ou type de modèle est le plus efficace pour détecter des anomalies dans les textes ? D'après les résultats expérimentaux, le modèle LLaMA 3 s'est révélé le plus efficace pour la détection d'anomalies textuelles, en offrant un bon équilibre entre le rappel et la précision. Ce modèle a particulièrement bien fonctionné dans les contextes où il est important de minimiser à la fois les faux positifs et les faux négatifs. D'autres modèles, tels que GPT-3.5 et Mistral, ont adopté une approche plus proactive, identifiant rapidement un large éventail d'anomalies. Cette agressivité peut être bénéfique dans des contextes où une détection exhaustive et rapide est essentielle, bien qu'elle entraîne une augmentation des faux positifs. Gemini Pro a montré une efficacité notable pour la détection d'anomalies thématiques, mais a rencontré des difficultés pour d'autres types d'anomalies. Ces résultats montrent que, bien que LLaMA 3 soit globalement le modèle le plus performant, le choix du modèle dépend largement du type d'anomalies à détecter et de l'exigence spécifique du scénario d'application.

Comment les techniques de conception de prompts influencent-elles les performances des LLMs ? La conception de prompts joue un rôle déterminant dans la performance des LLMs. Les prompts riches en exemples et les techniques de chaîne de pensée se sont révélées particulièrement efficaces pour guider les modèles dans des tâches complexes. Les résultats suggèrent que des prompts bien optimisés peuvent compenser certaines limites intrinsèques des modèles.

Quelles sont les limites des LLMs et comment pourraient-elles être surmontées ? Les limites des LLMs résident principalement dans leur difficulté à détecter des anomalies très subtiles ou spécifiques à un domaine. Des pistes d'amélioration incluent l'optimisation des prompts et l'adaptation des modèles à des contextes plus spécialisés. De futures recherches devront se concentrer sur ces ajustements pour maximiser l'efficacité des LLMs dans des environnements plus complexes.

* * *

Les résultats de cette troisième partie suggèrent que, bien que les LLMs aient montré des performances prometteuses, leur utilisation pour la détection d'anomalies textuelles peut encore être optimisée. Les axes de recherche futurs devraient porter sur l'amélioration des techniques de prompting, l'adaptation des modèles à des anomalies plus complexes et l'optimisation des ressources nécessaires pour traiter de grands volumes de données. De plus, l'intégration des LLMs dans des pipelines hybrides, combinant des approches traditionnelles et des techniques d'apprentissage profond, pourrait ouvrir de nouvelles perspectives pour une détection d'anomalies plus robuste et efficace.

Conclusion générale

La présente thèse a exploré les défis et les solutions associés à la détection d'anomalies textuelles dans le cadre de la veille stratégique. L'objectif principal était de développer et d'adapter des méthodes avancées de fouille de données et des grands modèles de langue (LLMs) pour améliorer la détection d'anomalies dans des données textuelles non structurées. À travers un examen approfondi de l'état de l'art, l'élaboration de méthodologies adaptées et des expérimentations empiriques, ce travail apporte des pistes de solutions spécifiques pour répondre aux problématiques posées.

* * *

La première partie de la thèse présente un panorama complet des approches de détection d'anomalies, avec un focus particulier sur les méthodes appliquées aux données textuelles. Le Chapitre 1 explore l'évolution des techniques de détection d'anomalies, en retraçant les progrès réalisés depuis les méthodes statistiques traditionnelles jusqu'aux approches modernes basées sur l'apprentissage automatique. Il met en lumière les tendances actuelles, telles que l'essor des réseaux neuronaux, l'apprentissage semi-supervisé, et les méthodes de calcul de scores innovantes. Le Chapitre 2 se concentre sur les défis uniques de la détection d'anomalies dans les textes, en présentant les techniques spécifiques adaptées à ce type de données non structurées. Il souligne l'importance des modèles de langue pré-entraînés, qui permettent une représentation plus efficace des données textuelles, et discute des applications potentielles dans le cadre de la veille stratégique. La partie conclut en identifiant les lacunes des approches actuelles et en définissant les axes de recherche qui seront développés dans la suite de la thèse.

La deuxième partie se concentre sur l'application et l'adaptation des techniques de fouille de données pour la détection d'anomalies dans les textes. Le Chapitre 3 détaille les méthodologies de pré-traitement et de représentation des données textuelles, en mettant l'accent sur l'importance de techniques avancées telles que Sentence-BERT. Il explore également l'intégration de divers algorithmes de détection, allant des méthodes traditionnelles aux approches incorporant des réseaux neuronaux. Le Chapitre 4 se concentre sur les ensembles de données utilisés pour les expérimentations, en décrivant les critères de sélection et les processus d'adaptation pour garantir une diversité et une pertinence des types d'anomalies. Le Chapitre 5 présente les résultats expérimentaux, démontrant comment les techniques de représentation avancées, combinées aux algorithmes d'apprentissage spécifiques, améliorent la détection d'anomalies textuelles, même avec un nombre limité de données annotées.

La troisième partie explore l'utilisation des Grands Modèles de Langue (LLMs) pour la détection d'anomalies textuelles. Le Chapitre 6 introduit les bases théoriques des LLMs et discute de leur architecture, leur échelle, et leurs capacités émergentes,

tout en soulignant leur pertinence pour la tâche de détection d'anomalies. Le Chapitre 7 développe une méthodologie basée sur l'utilisation des prompts pour maximiser la performance des LLMs dans des scénarios complexes. Ce chapitre se focalise sur la structuration des prompts et l'importance de la segmentation des données pour orienter efficacement les LLMs vers la détection d'anomalies. Le Chapitre 8 présente une série d'expérimentations qui évaluent la performance des LLMs en comparaison avec des méthodes de fouille de données classiques. Les résultats montrent que, dans des scénarios où les données annotées sont absentes, les LLMs surpassent les approches classiques. Cependant, leur avantage diminue dans des contextes où des données annotées sont disponibles, rendant les méthodes semi-supervisées ou faiblement supervisées plus efficaces. La partie conclut en proposant des perspectives d'intégration des LLMs dans les systèmes de veille stratégique, tout en explorant l'utilisation de ces modèles pour l'augmentation de données annotées et l'élaboration de systèmes hybrides combinant LLMs et méthodes traditionnelles pour une détection d'anomalies plus robuste et efficace.

* * *

À travers cette recherche, nous avons apporté des réponses concrètes aux questions de recherche posées en explorant les méthodologies de fouille de données et l'intégration des LLMs dans la détection d'anomalies textuelles.

Quels sont les principaux facteurs qui influencent l'efficacité des méthodes de fouille de données dans la détection d'anomalies pour les données textuelles non structurées ? Les résultats de cette thèse montrent que plusieurs facteurs influencent directement l'efficacité des méthodes de fouille de données dans la détection d'anomalies textuelles :

- **Techniques de représentation** : L'utilisation de techniques de représentation avancées, comme SBERT, améliore considérablement l'exactitude des modèles par rapport aux méthodes traditionnelles comme TF-IDF. Ces techniques permettent de capturer les relations sémantiques et le contexte des textes, facilitant ainsi la détection des anomalies complexes.
- **Niveau de supervision** : Les méthodes semi-supervisées PU et faiblement supervisées surpassent les méthodes non supervisées, même avec un nombre limité de données annotées. Ces techniques exploitent au mieux les exemples d'anomalies annotées (positifs), qui apportent plus d'informations que les exemples normaux.
- **Mécanismes de calcul de score** : Les mécanismes d'apprentissage de score de bout en bout montrent une performance supérieure aux méthodes traditionnelles à deux étapes pour le calcul des scores d'anomalie. Les méthodes basées sur l'apprentissage ensembliste, combinant plusieurs détecteurs, se révèlent également plus robustes.

Comment les LLMs se comparent-ils aux méthodes traditionnelles de fouille de données dans la détection d'anomalies textuelles ? Dominent-ils dans ce domaine comme dans d'autres ? Contrairement à d'autres domaines où les LLMs dominent, leur performance dans la détection d'anomalies textuelles est plus nuancée.

— **Scénarios sans données annotées** : Les LLMs se distinguent par leur capacité de *zero-shot* à détecter des anomalies, c'est-à-dire leur aptitude à effectuer des tâches sans être exposés à des exemples annotés spécifiques. Dans ces contextes, les LLMs surpassent largement les méthodes traditionnelles de fouille de données car ils sont capables d'adopter un cadre général de détection d'anomalies basé sur leur vaste pré-entraînement, ce qui les rend particulièrement efficaces pour détecter des comportements ou des patterns inhabituels même en l'absence totale de données annotées.
— **Scénarios avec données partiellement annotées** : Lorsque des données annotées sont partiellement disponibles, en particulier des exemples d'anomalies, les LLMs montrent des performances comparables aux méthodes semi-supervisées et faiblement supervisées. Toutefois, les méthodes de fouille de données peuvent surpasser nettement les LLMs dans certains cas, surtout lorsque les anomalies et les données normales sont moins distinguables. Cela s'explique par le fait que les LLMs, même avec des capacités de compréhension contextuelle, ne sont pas toujours optimisés pour ces distinctions fines sans un grand nombre d'exemples annotés.
— **Quantité d'exemples requis** : Les LLMs sont capables de fournir des résultats satisfaisants avec un nombre limité d'exemples annotés, ce qui les rend avantageux dans des contextes où les données annotées sont rares. Les méthodes de fouille de données, en revanche, nécessitent un plus grand nombre d'exemples pour atteindre une performance optimale, ce qui peut limiter leur efficacité lorsque les ressources en données annotées sont limitées.

Comment les contraintes pratiques, telles que les ressources computationnelles, le temps de traitement et la disponibilité des données, influencent-elles le choix des techniques de détection d'anomalies textuelles ? Les contraintes pratiques influencent fortement le choix des techniques à utiliser :

— **Ressources computationnelles et temps de traitement** : Les méthodes de fouille de données et les LLMs diffèrent considérablement en termes de besoins en ressources. Les méthodes de fouille de données peu profondes, comme XGBOD, sont rapides à entraîner et à exécuter, nécessitant peu de ressources, ce qui les rend adaptées aux applications en temps réel. En revanche, les modèles profonds et les LLMs demandent des infrastructures computationnelles plus importantes et un temps de traitement prolongé, limitant leur usage dans des environnements à ressources limitées ou nécessitant des réponses en temps réel.
— **Disponibilité des données annotées** : Les méthodes de fouille de données semi-supervisées et faiblement supervisées, telles que DevNet et XGBOD, offrent de meilleures performances lorsqu'un certain nombre d'exemples annotés est disponible. Elles exploitent efficacement les données annotées pour détecter des anomalies, mais leur efficacité diminue en l'absence de telles annotations. En revanche, les LLMs nécessitent moins d'exemples annotés pour fournir des résultats acceptables. Leur capacité à généraliser à partir d'un petit nombre d'exemples les rend particulièrement adaptés dans des scénarios où les annotations sont rares ou coûteuses à obtenir.

Quelles sont les perspectives de recherche pour améliorer la détection d'anomalies textuelles, notamment avec l'intégration des LLMs ? Les pers-

pectives de recherche pour améliorer la détection d'anomalies textuelles incluent :

— **Développement de modèles hybrides** : Combiner les LLMs avec des méthodes de fouille de données pour tirer parti des forces de chaque approche. Par exemple, les LLMs peuvent être utilisés pour la génération de données annotées synthétiques, qui peuvent ensuite être exploitées par des méthodes semi-supervisées et faiblement supervisées.

— **Ingénierie adaptative des prompts** : Concevoir des prompts dynamiques qui s'adaptent en fonction des caractéristiques des données d'entrée, permettant une amélioration continue des résultats en ajustant automatiquement les formulations des prompts selon les types d'anomalies détectées. Cela inclut des méthodes comme l'ajustement incrémental des prompts, où les prompts sont affinés au fur et à mesure que de nouvelles données deviennent disponibles.

— **Applications en temps réel et avec contraintes de ressources** : Explorer l'utilisation de LLMs légers ou de modèles optimisés, comme Gemini Nano, pour une détection d'anomalies efficace dans des environnements contraints en ressources computationnelles et en temps réel.

BIBLIOGRAPHIE

Abhaya, A. and Patra, B. K. (2023). An efficient method for autoencoder based outlier detection. *Expert Systems with Applications*, 213:118904. – Cité page 42.

Adikaram, K. K. L. B., Hussein, M. A., Effenberger, M., and Becker, T. (2015). Data Transformation Technique to Improve the Outlier Detection Power of Grubbs' Test for Data Expected to Follow Linear Relation. *Journal of Applied Mathematics*, 2015(1):708948. – Cité page 37.

Agarwal, S. and Sureka, A. (2015). Using KNN and SVM Based One-Class Classifier for Detecting Online Radicalization on Twitter. In Natarajan, R., Barua, G., and Patra, M. R., editors, *Distributed Computing and Internet Technology*, volume 8956, pages 431–442. Springer International Publishing, Cham. – Cité pages 31, 39 et 60.

Aggarwal, C. C. (2017a). *Outlier Analysis*. Springer International Publishing, Cham. – Cité pages 21, 28, 30, 31, 32, 37, 39, 40, 42 et 121.

Aggarwal, C. C. (2017b). Proximity-Based Outlier Detection. In Aggarwal, C. C., editor, *Outlier Analysis*, pages 111–147. Springer International Publishing, Cham. – Cité page 39.

Aggarwal, C. C. and Reddy, C. K., editors (2018). *Data Clustering: Algorithms and Applications*. Chapman and Hall/CRC, 1 edition. – Cité page 21.

Agrawal, S. and Agrawal, J. (2015). Survey on Anomaly Detection using Data Mining Techniques. *Procedia Computer Science*, 60:708–713. – Cité page 37.

Ahmed, M., Choudhury, N., and Uddin, S. (2017). Anomaly Detection on Big Data in Financial Markets. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 998–1001, Sydney Australia. ACM. – Cité page 20.

Ahmed, M., Mahmood, A. N., and Islam, M. R. (2016a). A survey of anomaly detection techniques in financial domain. *Future Generation Computer Systems*, 55:278–288. – Cité pages 20 et 21.

Ahmed, M., Naser Mahmood, A., and Hu, J. (2016b). A survey of network anomaly detection techniques. *Journal of Network and Computer Applications*, 60:19–31. – Cité page 43.

Ait-Saada, M. and Nadif, M. (2023). Unsupervised Anomaly Detection in Multi-Topic Short-Text Corpora. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1392–1403, Dubrovnik, Croatia. Association for Computational Linguistics. – Cité pages 21, 39, 68, 71 et 113.

Akcay, S., Atapour-Abarghouei, A., and Breckon, T. P. (2019). GANomaly: Semi-supervised Anomaly Detection via Adversarial Training. In Jawahar, C. V., Li, H., Mori, G., and Schindler, K., editors, *Computer Vision – ACCV 2018*, pages 622–637, Cham. Springer International Publishing. – Cité pages 31, 32 et 43.

Al-amri, R., Murugesan, R. K., Man, M., Abdulateef, A. F., Al-Sharafi, M. A., and Alkahtani, A. A. (2021). A Review of Machine Learning and Deep Learning Techniques for Anomaly Detection in IoT Data. *Applied Sciences*, 11(12):5320. – Cité pages 20 et 37.

Al-taei, R. and Haeri, M. A. (2019). An Ensemble Angle-Based Outlier Detection for Big Data. In Grandinetti, L., Mirtaheri, S. L., and Shahbazian, R., editors, *High-Performance Computing and Big Data Analysis*, pages 98–108, Cham. Springer International Publishing. – Cité page 40.

Aldera, S., Emam, A., Al-Qurishi, M., Alrubaian, M., and Alothaim, A. (2021). Exploratory Data Analysis and Classification of a New Arabic Online Extremism Dataset. *IEEE Access*, 9:161613–161626. – Cité page 60.

Aleskerov, E., Freisleben, B., and Rao, B. (1997). CARDWATCH: A neural network based database mining system for credit card fraud detection. In *Proceedings of the IEEE/IAFE 1997 Computational Intelligence for Financial Engineering (CIFEr)*, pages 220–226. – Cité page 26.

Ali, M. Q., Al-Shaer, E., Khan, H., and Khayam, S. A. (2013). Automated Anomaly Detector Adaptation using Adaptive Threshold Tuning. *ACM Trans. Inf. Syst. Secur.*, 15(4):17:1–17:30. – Cité page 26.

Allan, J., Carbonell, J. G., Doddington, G., Yamron, J., and Yang, Y. (1998). Topic detection and tracking pilot study final report. – Cité page 64.

Allan, J., Lavrenko, V., and Jin, H. (2000). First story detection in TDT is hard. In *Proceedings of the Ninth International Conference on Information and Knowledge Management - CIKM '00*, pages 374–381, McLean, Virginia, United States. ACM Press. – Cité pages 64, 108 et 109.

Aminikhanghahi, S. and Cook, D. J. (2017). A survey of methods for time series change point detection. *Knowledge and Information Systems*, 51(2):339–367. – Cité pages 21 et 52.

Amiri, F., Rezaei Yousefi, M., Lucas, C., Shakery, A., and Yazdani, N. (2011). Mutual information-based feature selection for intrusion detection systems. *Journal of Network and Computer Applications*, 34(4):1184–1199. – Cité page 44.

An, J. and Cho, S. (2015). Variational autoencoder based anomaly detection using reconstruction probability. *Special Lecture on IE*, 2(1):1–18. – Cité page 42.

An, P., Wang, Z., and Zhang, C. (2022). Ensemble unsupervised autoencoders and Gaussian mixture model for cyberattack detection. *Information Processing & Management*, 59(2):102844. – Cité page 39.

Anderson, A. and Haas, H. (2011). Kullback-Leibler Divergence (KLD) Based Anomaly Detection and Monotonic Sequence Analysis. In *2011 IEEE Vehicular Technology Conference (VTC Fall)*, pages 1–5. – Cité page 44.

Ankerst, M., Breunig, M. M., Kriegel, H.-P., and Sander, J. (1999). OPTICS: Ordering points to identify the clustering structure. *ACM Sigmod record*, 28(2):49–60. – Cité page 40.

Anscombe, F. J. and Guttman, I. (1960). Rejection of Outliers. *Technometrics*, 2(2):123–147. – Cité page 26.

Arning, A., Agrawal, R., and Raghavan, P. (1996). A linear method for deviation detection in large databases. In *KDD*, volume 1141, pages 972–981. – Cité page 21.

Atluri, G., Karpatne, A., and Kumar, V. (2018). Spatio-Temporal Data Mining: A Survey of Problems and Methods. *ACM Comput. Surv.*, 51(4):83:1–83:41. – Cité page 53.

Attar, A. E., Khatoun, R., and Lemercier, M. (2014). A Gaussian mixture model for dynamic detection of abnormal behavior in smartphone applications. In *2014 Global Information Infrastructure and Networking Symposium (GIIS)*, pages 1–6. – Cité page 38.

Avola, D., Cinque, L., Di Mambro, A., Diko, A., Fagioli, A., Foresti, G. L., Marini, M. R., Mecca, A., and Pannone, D. (2022). Low-Altitude Aerial Video Surveillance via One-Class SVM Anomaly Detection from Textural Features in UAV Images. *Information*, 13(1):2. – Cité page 41.

Bahrololum, M. and Khaleghi, M. (2008). Anomaly Intrusion Detection System Using Gaussian Mixture Model. In *2008 Third International Conference on Convergence and Hybrid Information Technology*, volume 1, pages 1162–1167. – Cité page 38.

BalaAnand, M., Karthikeyan, N., Karthik, S., Varatharajan, R., Manogaran, G., and Sivaparthipan, C. B. (2019). An enhanced graph-based semi-supervised learning algorithm to detect fake users on Twitter. *The Journal of Supercomputing*, 75(9):6085–6105. – Cité page 54.

Banerjee, A., Burlina, P., and Meth, R. (2007). Fast Hyperspectral Anomaly Detection via SVDD. In *2007 IEEE International Conference on Image Processing*, volume 4, pages IV – 101–IV – 104. – Cité page 41.

Barrett, L., Fletcher, S., and Kingan, R. (2019). Textual Outlier Detection and Anomalies in Financial Reporting. In *2nd KDD Workshop on Anomaly Detection in Finance*, page 6. – Cité pages 21, 40, 60, 66, 67 et 107.

Baur, C., Denner, S., Wiestler, B., Navab, N., and Albarqouni, S. (2021). Autoencoders for unsupervised anomaly segmentation in brain MR images: A comparative study. *Medical Image Analysis*, 69:101952. – Cité page 53.

Bayarri, MJ. and Morales, J. (2003). Bayesian measures of surprise for outlier detection. *Journal of Statistical Planning and Inference*, 111(1-2):3–22. – Cité page 21.

Beghi, A., Cecchinato, L., Corazzol, C., Rampazzo, M., Simmini, F., and Susto, G. (2014). A One-Class SVM Based Tool for Machine Learning Novelty Detection in HVAC Chiller Systems. *IFAC Proceedings Volumes*, 47(3):1953–1958. – Cité page 41.

Bejan, M., Manolache, A., and Popescu, M. (2023). AD-NLP: A Benchmark for Anomaly Detection in Natural Language Processing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10766–10778, Singapore. Association for Computational Linguistics. – Cité pages 21, 66, 107 et 108.

Bekker, J. and Davis, J. (2020). Learning from positive and unlabeled data: A survey. *Machine Learning*, 109(4):719–760. – Cité page 32.

Belton, N., Hagos, M. T., Lawlor, A., and Curran, K. M. (2023). Fewsome: One-class few shot anomaly detection with siamese networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2978–2987. – Cité page 32.

Bereziński, P., Jasiul, B., and Szpyrka, M. (2015). An Entropy-Based Network Anomaly Detection Method. *Entropy*, 17(4):2367–2408. – Cité page 44.

Beula Rani, B. J. and Sumathi M. E, L. (2020). Survey on Applying GAN for Anomaly Detection. In *2020 International Conference on Computer Communication and Informatics (ICCCI)*, pages 1–5. – Cité page 43.

Bian, J., Hui, X., Sun, S., Zhao, X., and Tan, M. (2019). A Novel and Efficient CVAE-GAN-Based Approach With Informative Manifold for Semi-Supervised Anomaly Detection. *IEEE Access*, 7:88903–88916. – Cité page 44.

Biswas, P. and Samanta, T. (2021). Anomaly detection using ensemble random forest in wireless sensor network. *International Journal of Information Technology*, 13(5):2043–2052. – Cité page 45.

Bobur, M., Aibek, K., Abay, B., and Hajiyev, F. (2020). Anomaly Detection Between Judicial Text-Based Documents. In *2020 IEEE 14th International Conference on Application of Information and Communication Technologies (AICT)*, pages 1–5. – Cité page 59.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146. – Cité page 68.

Borne, K. D. and Vedachalam, A. (2012). Surprise detection in multivariate astronomical data. In *Statistical Challenges in Modern Astronomy V*, pages 275–289. Springer. – Cité page 21.

Boukela, L., Zhang, G., Yacoub, M., Bouzefrane, S., Ahmadi, S. B. B., and Jelodar, H. (2021). A modified LOF-based approach for outlier characterization in IoT. *Annals of Telecommunications*, 76(3):145–153. – Cité pages 35 et 40.

Boukerche, A., Zheng, L., and Alfandi, O. (2021). Outlier Detection: Methods, Models, and Classification. *ACM Computing Surveys*, 53(3):1–37. – Cité page 21.

Bounoua, W., Benkara, A. B., Kouadri, A., and Bakdi, A. (2020). Online monitoring scheme using principal component analysis through Kullback-Leibler divergence analysis technique for fault detection. *Transactions of the Institute of Measurement and Control*, 42(6):1225–1238. – Cité page 44.

Boutalbi, K., Loukil, F., Verjus, H., Telisson, D., and Salamatian, K. (2023). Machine learning for text anomaly detection: A systematic review. In *The 5th IEEE International Workshop on Deep Analysis of Data-Driven Applications*, Turin, Italy. – Cité page 72.

Box, G. E. P. and Tiao, G. C. (1968). A Bayesian approach to some outlier problems. *Biometrika*, 55(1):119–129. – Cité page 26.

Brauckhoff, D., Salamatian, K., and May, M. (2009). Applying PCA for Traffic Anomaly Detection: Problems and Solutions. In *IEEE INFOCOM 2009 - The 28th Conference on Computer Communications*, pages 2866–2870, Rio De Janeiro, Brazil. IEEE. – Cité pages 35 et 42.

Brause, R., Langsdorf, T., and Hepp, M. (1999). Neural data mining for credit card fraud detection. In *Proceedings 11th International Conference on Tools with Artificial Intelligence*, pages 103–106. – Cité page 26.

Breidenstein, A. and Labeau, M. (2024). Using Locally Learnt Word Representations for better Textual Anomaly Detection. In Tafreshi, S., Akula, A., Sedoc, J., Drozd, A., Rogers, A., and Rumshisky, A., editors, *Proceedings of the Fifth Workshop on Insights from Negative Results in NLP*, pages 82–91, Mexico City, Mexico. Association for Computational Linguistics. – Cité page 66.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. (1999). OPTICS-OF: Identifying Local Outliers. In Żytkow, J. M. and Rauch, J., editors, *Principles of Data Mining and Knowledge Discovery*, pages 262–270, Berlin, Heidelberg. Springer. – Cité page 40.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. (2000). LOF: Identifying density-based local outliers. *ACM SIGMOD Record*, 29(2):93–104. – Cité pages 35, 40 et 122.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language Models are Few-Shot Learners. *Advances in neural information processing systems*, 33:1877–1901. – Cité pages 164, 169 et 172.

Callegari, C., Gazzarrini, L., Giordano, S., Pagano, M., and Pepe, T. (2014). Improving PCA-based anomaly detection by using multiple time scale analysis and Kullback–Leibler divergence. *International Journal of Communication Systems*, 27(10):1731–1751. – Cité page 44.

Cao, Y., Li, Y., Coleman, S., Belatreche, A., and McGinnity, T. M. (2015). Adaptive Hidden Markov Model With Anomaly States for Price Manipulation Detection. *IEEE Transactions on Neural Networks and Learning Systems*, 26(2):318–330. – Cité page 52.

Carrasco, J., López, D., Aguilera-Martos, I., García-Gil, D., Markova, I., García-Barzana, M., Arias-Rodil, M., Luengo, J., and Herrera, F. (2021). Anomaly detection in predictive maintenance: A new evaluation framework for temporal un-

supervised anomaly detection algorithms. *Neurocomputing*, 462:440–452. – Cité page 52.

Castro Gertrudes, J., Zimek, A., Sander, J., and Campello, R. J. G. B. (2019). A unified view of density-based methods forsemi-supervised clustering and classification. *Data Mining and Knowledge Discovery*, 33(6):1894–1952. – Cité page 32.

Çelik, M., Dadaşer-Çelik, F., and Dokuz, A. Ş. (2011). Anomaly detection in temperature data using DBSCAN algorithm. In *2011 International Symposium on Innovations in Intelligent Systems and Applications*, pages 91–95. IEEE. – Cité page 40.

Chahla, C., Snoussi, H., Merghem, L., and Esseghir, M. (2020). A deep learning approach for anomaly detection and prediction in power consumption data. *Energy Efficiency*, 13(8):1633–1651. – Cité page 52.

Chalapathy, R. and Chawla, S. (2019). Deep Learning for Anomaly Detection: A Survey. *arXiv:1901.03407 [cs, stat]*. – Cité pages 27 et 37.

Chalapathy, R., Menon, A. K., and Chawla, S. (2019). Anomaly Detection using One-Class Neural Networks. *arXiv:1802.06360 [cs, stat]*. – Cité pages 21 et 36.

Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):1–58. – Cité pages 21, 27, 28, 30, 31, 32, 37, 39, 43 et 52.

Chauhan, P. and Shukla, M. (2015). A review on outlier detection techniques on data stream by using different approaches of K-Means algorithm. In *2015 International Conference on Advances in Computer Engineering and Applications*, pages 580–585. IEEE. – Cité page 35.

Chauhan, S. and Vig, L. (2015). Anomaly detection in ECG time signals via deep long short-term memory networks. In *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–7. – Cité page 52.

Chen, D., Lu, C.-T., Kou, Y., and Chen, F. (2008). On Detecting Spatial Outliers. *GeoInformatica*, 12(4):455–475. – Cité page 53.

Chen, J., Sathe, S., Aggarwal, C., and Turaga, D. (2017). Outlier Detection with Autoencoder Ensembles. In *Proceedings of the 2017 SIAM International Conference on Data Mining (SDM)*, Proceedings, pages 90–98. Society for Industrial and Applied Mathematics. – Cité page 42.

Chen, L., Wang, W., and Yang, Y. (2021a). CELOF: Effective and fast memory efficient local outlier detection in high-dimensional data streams. *Applied Soft Computing*, 102:107079. – Cité page 40.

Chen, L.-J., Ho, Y.-H., Hsieh, H.-H., Huang, S.-T., Lee, H.-C., and Mahajan, S. (2018a). ADF: An Anomaly Detection Framework for Large-Scale PM2.5 Sensing Systems. *IEEE Internet of Things Journal*, 5(2):559–570. – Cité page 53.

Chen, T. and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. Association for Computing Machinery. – Cité page 45.

Chen, W., Zhang, Y., Yeo, C. K., Lau, C. T., and Lee, B. S. (2018b). Unsupervised rumor detection based on users' behaviors using neural networks. *Pattern Recognition Letters*, 105:226–233. – Cité page 60.

Chen, Y., Miao, D., and Zhang, H. (2010). Neighborhood outlier detection. *Expert Systems with Applications*, 37(12):8745–8749. – Cité page 39.

Chen, Y., Zhao, Q., and Lu, L. (2021b). Combining the outputs of various k-nearest neighbor anomaly detectors to form a robust ensemble model for high-dimensional geochemical anomaly detection. *Journal of Geochemical Exploration*, 231:106875. – Cité page 39.

Chen, Z., Yeo, C. K., Lee, B. S., and Lau, C. T. (2018c). Autoencoder-based network anomaly detection. In *2018 Wireless Telecommunications Symposium (WTS)*, pages 1–5, Phoenix, AZ. IEEE. – Cité page 42.

Cheon, S.-P., Kim, S., Lee, S.-Y., and Lee, C.-B. (2009). Bayesian networks based rare event prediction with sensor data. *Knowledge-Based Systems*, 22(5):336–343. – Cité page 21.

Cheong, M.-S., Wu, M.-C., and Huang, S.-H. (2021). Interpretable Stock Anomaly Detection Based on Spatio-Temporal Relation Networks With Genetic Algorithm. *IEEE Access*, 9:68302–68319. – Cité page 20.

Chou, J.-S. and Telaga, A. S. (2014). Real-time detection of anomalous power consumption. *Renewable and Sustainable Energy Reviews*, 33:400–411. – Cité page 52.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A. M., Pillai, T. S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., and Fiedel, N. (2022). PaLM: Scaling Language Modeling with Pathways. – Cité page 164.

Christophe, C., Velcin, J., Cugliari, J., Suignard, P., and Boumghar, M. (2020). How to Detect Novelty in Textual Data Streams? A Comparative Study of Existing Methods. In Lemaire, V., Malinowski, S., Bagnall, A., Bondu, A., Guyet, T., and Tavenard, R., editors, *Advanced Analytics and Learning on Temporal Data*, Lecture Notes in Computer Science, pages 110–125, Cham. Springer International Publishing. – Cité page 65.

Chuah, M. C. and Fu, F. (2007). ECG Anomaly Detection via Time Series Analysis. In Thulasiraman, P., He, X., Xu, T. L., Denko, M. K., Thulasiram, R. K., and Yang, L. T., editors, *Frontiers of High Performance Computing and Networking ISPA 2007 Workshops*, pages 123–135, Berlin, Heidelberg. Springer. – Cité page 52.

Cichosz, P. (2020). Unsupervised modeling anomaly detection in discussion forums posts using global vectors for text representation. *Natural Language Engineering*, 26(5):551–578. – Cité page 68.

Cieri, C., Graff, D., Liberman, M., Martey, N., Strassel, S., et al. (1999a). The TDT-2 text and speech corpus. In *Proceedings of the DARPA Broadcast News Workshop*, pages 57–60. – Cité page 64.

Cieri, C., Graff, D., Liberman, M., Martey, N., Strassel, S., et al. (1999b). The TDT-2 text and speech corpus. In *Proceedings of the DARPA Broadcast News Workshop*, pages 57–60. – Cité page 108.

Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020). ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. – Cité pages 70 et 71.

Clifton, D. A., Hugueny, S., and Tarassenko, L. (2009). A comparison of approaches to multivariate extreme value theory for novelty detection. In *2009 IEEE/SP 15th Workshop on Statistical Signal Processing*, pages 13–16, Cardiff, United Kingdom. IEEE. – Cité page 37.

Coelho, D., Costa, D., Rocha, E. M., Almeida, D., and Santos, J. P. (2022). Predictive maintenance on sensorized stamping presses by time series segmentation, anomaly detection, and classification algorithms. *Procedia Computer Science*, 200:1184–1193. – Cité page 52.

Cook, A. A., Mısırlı, G., and Fan, Z. (2020). Anomaly Detection for IoT Time-Series Data: A Survey. *IEEE Internet of Things Journal*, 7(7):6481–6494. – Cité page 20.

Cortal, G. (2022). Covid-19-french news dataset. – Cité page 108.

Cuina, L., Tianqi, L., and Dongli, W. (2019). Study on Anomaly Data Detection Method for Automatic Soil Moisture Observation. In *2019 International Conference on Meteorology Observations (ICMO)*, pages 1–4. – Cité page 53.

Daneshpazhouh, A. and Sami, A. (2015). Semi-Supervised Outlier Detection with Only Positive and Unlabeled Data Based on Fuzzy Clustering. *International Journal on Artificial Intelligence Tools*, 24(03):1550003. – Cité page 31.

Dang, T. T., Ngan, H. Y., and Liu, W. (2015). Distance-based k-nearest neighbors outlier detection method in large-scale traffic data. In *2015 IEEE International Conference on Digital Signal Processing (DSP)*, pages 507–510, Singapore, Singapore. IEEE. – Cité pages 31 et 39.

Das, A. S., Ajay, A., Saha, S., and Bhuyan, M. (2024). Few-Shot Anomaly Detection in Text with Deviation Learning. In Luo, B., Cheng, L., Wu, Z.-G., Li, H., and Li, C., editors, *Neural Information Processing*, pages 425–438, Singapore. Springer Nature. – Cité pages 66, 70 et 71.

Das, M. and Parthasarathy, S. (2009). Anomaly detection and spatio-temporal analysis of global climate system. In *Proceedings of the Third International Workshop on Knowledge Discovery from Sensor Data*, SensorKDD '09, pages 142–150, New York, NY, USA. Association for Computing Machinery. – Cité page 52.

Dasigi, P. and Hovy, E. (2014). Modeling Newswire Events using Neural Networks for Anomaly Detection. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, page 9. – Cité pages 60 et 65.

Davies, D. L. and Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227. – Cité page 47.

De Benedetti, M., Leonardi, F., Messina, F., Santoro, C., and Vasilakos, A. (2018). Anomaly detection and predictive maintenance for photovoltaic systems. *Neurocomputing*, 310:59–68. – Cité page 52.

de la Torre-Abaitua, G., Lago-Fernández, L. F., and Arroyo, D. (2021). A Compression-Based Method for Detecting Anomalies in Textual Data. *Entropy*, 23(5):618. – Cité page 66.

De Melo Borges, J., Riedel, T., and Beigl, M. (2016). Urban Anomaly Detection: A Use-Case for Participatory Infra-Structure Monitoring. In *Proceedings of the Second International Conference on IoT in Urban Space*, Urb-IoT '16, pages 36–38, New York, NY, USA. Association for Computing Machinery. – Cité page 53.

Deng, J. and Brown, E. T. (2022). SSDBCODI: Semi-Supervised Density-Based Clustering with Outliers Detection Integrated. – Cité page 32.

Deng, J., Zhou, J., Sun, H., Zheng, C., Mi, F., Meng, H., and Huang, M. (2022). COLD: A benchmark for chinese offensive language detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11580–11599. – Cité page 112.

Denning, D. (1987). An Intrusion-Detection Model. *IEEE Transactions on Software Engineering*, SE-13(2):222–232. – Cité page 26.

Derhab, A., Belaoued, M., Mohiuddin, I., Kurniawan, F., and Khan, M. K. (2022). Histogram-Based Intrusion Detection and Filtering Framework for Secure and Safe In-Vehicle Networks. *IEEE Transactions on Intelligent Transportation Systems*, 23(3):2366–2379. – Cité page 38.

Devaki, R., Kathiresan, V., and Gunasekaran, S. (2014). Credit card fraud detection using time series analysis. *International Journal of Computer Applications*, 3:8–10. – Cité page 52.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*. – Cité pages 70, 82, 84 et 159.

Di Biase, G., Blum, H., Siegwart, R., and Cadena, C. (2021). Pixel-Wise Anomaly Detection in Complex Driving Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16918–16927. – Cité page 54.

Di Mattia, F., Galeone, P., De Simoni, M., and Ghelfi, E. (2021). A Survey on GANs for Anomaly Detection. – Cité page 43.

Ding, H., Huang, N., Wu, Y., and Cui, X. (2024). LEGAN: Addressing Intraclass Imbalance in GAN-Based Medical Image Augmentation for Improved Imbalanced Data Classification. *IEEE Transactions on Instrumentation and Measurement*, 73:1–14. – Cité page 40.

Ding, K., Li, J., Agarwal, N., and Liu, H. (2021). Inductive anomaly detection on attributed networks. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20, pages 1288–1294, Yokohama, Yokohama, Japan. – Cité page 30.

Ding, K., Li, J., Bhanushali, R., and Liu, H. (2019). Deep anomaly detection on attributed networks. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pages 594–602. SIAM. – Cité pages 27 et 34.

Ding, Q. and Kolaczyk, E. D. (2013). A Compressed PCA Subspace Method for Anomaly Detection in High-Dimensional Data. *IEEE Transactions on Information Theory*, 59(11):7419–7433. – Cité page 42.

Ding, Z. and Fei, M. (2013). An Anomaly Detection Approach Based on Isolation Forest Algorithm for Streaming Data using Sliding Window. *IFAC Proceedings Volumes*, 46(20):12–17. – Cité page 45.

Domingues, R., Filippone, M., Michiardi, P., and Zouaoui, J. (2018). A comparative evaluation of outlier detection algorithms: Experiments and analyses. *Pattern Recognition*, 74:406–421. – Cité page 21.

Donahue, J., Krähenbühl, P., and Darrell, T. (2017). Adversarial Feature Learning. – Cité page 43.

Dong, X., Yu, Z., Cao, W., Shi, Y., and Ma, Q. (2020). A survey on ensemble learning. *Frontiers of Computer Science*, 14:241–258. – Cité page 45.

Donoho, S. (2004). Early detection of insider trading in option markets. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 420–429, New York, NY, USA. Association for Computing Machinery. – Cité page 26.

Dorbu, F. E. and Hashemi-Beni, L. (2023). Geospatial Intelligence for Individual Crop Detection and Anomaly Monitoring. In *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, pages 1051–1054. – Cité page 53.

Duan, L., Xu, L., Liu, Y., and Lee, J. (2009). Cluster-based outlier detection. *Annals of Operations Research*, 168(1):151–168. – Cité pages 39 et 40.

Duong, P., Nguyen, V., Dinh, M., Le, T., Tran, D., and Ma, W. (2015). Graph-based semi-supervised Support Vector Data Description for novelty detection. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6. – Cité pages 31 et 41.

Eberle, W., Graves, J., and Holder, L. (2010). Insider Threat Detection Using a Graph-Based Approach. *Journal of Applied Security Research*, 6(1):32–81. – Cité page 54.

Eberle, W. and Holder, L. (2009). Graph-based approaches to insider threat detection. In *Proceedings of the 5th Annual Workshop on Cyber Security and Information Intelligence Research: Cyber Security and Information Intelligence Challenges and Strategies*, CSIIRW '09, pages 1–4, New York, NY, USA. Association for Computing Machinery. – Cité page 54.

Edgeworth, F. (1887). XLI. *On discordant observations*. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 23(143):364–375. – Cité page 26.

Emin Sahin, M. (2022). Deep learning-based approach for detecting COVID-19 in chest X-rays. *Biomedical Signal Processing and Control*, 78:103977. – Cité page 53.

Erfani, M., Shoeleh, F., and Ghorbani, A. A. (2020). Financial Fraud Detection using Deep Support Vector Data Description. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 2274–2282. – Cité page 41.

Eskin, E. (2000). Detecting Errors within a Corpus using Anomaly Detection. In *1st Meeting of the North American Chapter of the Association for Computational Linguistics*. – Cité page 60.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). Density-based spatial clustering of applications with noise. In *Int. Conf. Knowledge Discovery and Data Mining*, volume 240. – Cité page 40.

Fahim, M. and Sillitti, A. (2019). Anomaly Detection, Analysis and Prediction Techniques in IoT Environment: A Systematic Literature Review. *IEEE Access*, 7:81664–81681. – Cité page 20.

Fan, Z., Yin, J., Song, Y., and Liu, Z. (2020). Real-time and accurate abnormal behavior detection in videos. *Machine Vision and Applications*, 31(7):72. – Cité page 54.

Fatemifar, S., Awais, M., Akbari, A., and Kittler, J. (2020). A Stacking Ensemble for Anomaly Based Client-Specific Face Spoofing Detection. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 1371–1375. – Cité page 45.

Fearnhead, P. and Rigaill, G. (2019). Changepoint detection in the presence of outliers. *Journal of the American Statistical Association*, 114(525):169–183. – Cité page 21.

Ferret, O. (2021). Exploration des relations sémantiques sous-jacentes aux plongements contextuels de mots. In *Traitement Automatique Des Langues Naturelles*, pages 26–36. ATALA. – Cité page 85.

Fiscus, J., Doddington, G., Garofolo, J., and Martin, A. (1999). NIST's 1998 Topic Detection and Tracking evaluation (TDT2). In *Proceedings of the 1999 DARPA Broadcast News Workshop*, pages 19–24. – Cité pages 64 et 108.

Foorthuis, R. (2021). On the nature and types of anomalies: A review of deviations in data. *International Journal of Data Science and Analytics*, 12(4):297–331. – Cité page 21.

Fouche, E., Meng, Y., Guo, F., Zhuang, H., Bohm, K., and Han, J. (2020). Mining Text Outliers in Document Directories. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 152–161, Sorrento, Italy. IEEE. – Cité page 61.

Fox, A. J. (1972). Outliers in Time Series. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(3):350–363. – Cité page 26.

Frikha, A., Krompaß, D., Köpken, H.-G., and Tresp, V. (2021). Few-Shot One-Class Classification via Meta-Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(8):7448–7456. – Cité page 32.

Galvão, Y. M., Albuquerque, V. A., Fernandes, B. J. T., and Valença, M. J. S. (2017). Anomaly detection in smart houses: Monitoring elderly daily behavior for fall detecting. In *2017 IEEE Latin American Conference on Computational Intelligence (LA-CCI)*, pages 1–6. – Cité page 54.

Gao, J., Hu, W., Zhang, Z. M., Zhang, X., and Wu, O. (2011). RKOF: Robust Kernel-Based Local Outlier Detection. In Huang, J. Z., Cao, L., and Srivastava, J., editors, *Advances in Knowledge Discovery and Data Mining*, pages 270–283, Berlin, Heidelberg. Springer. – Cité page 38.

Garcia, G., Afonso, L., Passos, L., Jodas, D., P. Da Costa, K., and Papa, J. (2023). FakeRecogna Anomaly: Fake News Detection in a New Brazilian Corpus. In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pages 830–837, Lisbon, Portugal. SCITEPRESS - Science and Technology Publications. – Cité page 60.

Garcia-Teodoro, P., Diaz-Verdejo, J., Maciá-Fernández, G., and Vázquez, E. (2009). Anomaly-based network intrusion detection: Techniques, systems and challenges. *computers & security*, 28(1-2):18–28. – Cité page 20.

Ghosal, T., Saikh, T., Biswas, T., Ekbal, A., and Bhattacharyya, P. (2022). Novelty Detection: A Perspective from Natural Language Processing. *Computational Linguistics*, 48(1):77–117. – Cité pages 64 et 108.

Ghosal, T., Salam, A., Tiwari, S., Ekbal, A., and Bhattacharyya, P. (2018). TAP-DLND 1.0 : A Corpus for Document Level Novelty Detection. page 7. – Cité page 64.

Ghosh, A., Wanken, J., and Charron, F. (1998). Detecting anomalous and unknown intrusions against programs. In *Proceedings 14th Annual Computer Security Applications Conference (Cat. No.98EX217)*, pages 259–267. – Cité page 26.

Goernitz, N., Kloft, M., Rieck, K., and Brefeld, U. (2013). Toward Supervised Anomaly Detection. *Journal of Artificial Intelligence Research*, 46:235–262. – Cité pages 28, 29 et 30.

Gogoi, P., Bhattacharyya, D. K., Borah, B., and Kalita, J. K. (2011). A Survey of Outlier Detection Methods in Network Anomaly Identification. *The Computer Journal*, 54(4):570–588. – Cité page 20.

Goldstein, M. and Dengel, A. (2012). Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. *KI-2012: poster and demo track*, 1:59–63. – Cité pages 35, 38 et 122.

Goldstein, M. and Uchida, S. (2016). A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data. *PLOS ONE*, 11(4):e0152173. – Cité page 31.

Golmohammadi, K. and Zaiane, O. R. (2015). Time series contextual anomaly detection for detecting market manipulation in stock market. In *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10. – Cité pages 20 et 52.

Golmohammadi, K. and Zaiane, O. R. (2017). Sentiment Analysis on Twitter to Improve Time Series Contextual Anomaly Detection for Detecting Stock Market Manipulation. In Bellatreche, L. and Chakravarthy, S., editors, *Big Data Analytics and Knowledge Discovery*, Lecture Notes in Computer Science, pages 327–342, Cham. Springer International Publishing. – Cité page 52.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Networks. *arXiv:1406.2661 [cs, stat]*. – Cité page 43.

Graff, D., Cieri, C., Strassel, S., and Martey, N. (1999). The TDT-3 text and speech corpus. In *Proceedings of DARPA Broadcast News Workshop*, pages 57–60. – Cité page 64.

Greifeneder, F., Khamala, E., Sendabo, D., Wagner, W., Zebisch, M., Farah, H., and Notarnicola, C. (2019). Detection of soil moisture anomalies based on Sentinel-1. *Physics and Chemistry of the Earth, Parts A/B/C*, 112:75–82. – Cité page 53.

Gröndahl, T., Pajola, L., Juuti, M., Conti, M., and Asokan, N. (2018). All You Need is "Love": Evading Hate-speech Detection. – Cité page 60.

Grubbs, F. E. (1969). Procedures for Detecting Outlying Observations in Samples. *Technometrics*, 11(1):1–21. – Cité pages 26 et 37.

Gu, X., Akoglu, L., and Rinaldo, A. (2019). Statistical Analysis of Nearest Neighbor Methods for Anomaly Detection. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc. – Cité pages 30, 39 et 64.

Guille, A. and Favre, C. (2015). Event detection, tracking, and visualization in Twitter: A mention-anomaly-based approach. *Social Network Analysis and Mining*, 5(1):18. – Cité page 60.

Gupta, D., Gupta, M., Bhatt, S., and Tosun, A. S. (2021). Detecting Anomalous User Behavior in Remote Patient Monitoring. In *2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 33–40. – Cité page 20.

Gutflaish, E., Kontorovich, A., Sabato, S., Biller, O., and Sofer, O. (2019). Temporal anomaly detection: Calibrating the surprise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3755–3762. – Cité page 21.

Hamdi, S., Bouindour, S., Snoussi, H., Wang, T., and Abid, M. (2021). End-to-End Deep One-Class Learning for Anomaly Detection in UAV Video Stream. *Journal of Imaging*, 7(5):90. – Cité page 46.

Hamunyela, E., Brandt, P., Shirima, D., Do, H. T. T., Herold, M., and Roman-Cuesta, R. M. (2020). Space-time detection of deforestation, forest degradation and regeneration in montane forests of Eastern Tanzania. *International Journal of Applied Earth Observation and Geoinformation*, 88:102063. – Cité page 53.

Hamunyela, E., Verbesselt, J., De Bruin, S., and Herold, M. (2016). Monitoring Deforestation at Sub-Annual Scales as Extreme Events in Landsat Data Cubes. *Remote Sensing*, 8(8):651. – Cité page 53.

Han, C., Rundo, L., Murao, K., Noguchi, T., Shimahara, Y., Milacski, Z. Á., Koshino, S., Sala, E., Nakayama, H., and Satoh, S. (2021a). MADGAN: Unsupervised medical anomaly detection GAN using multiple adjacent brain MRI slice reconstruction. *BMC Bioinformatics*, 22(2):31. – Cité page 53.

Han, S., Hu, X., Huang, H., Jiang, M., and Zhao, Y. (2022). ADBench: Anomaly Detection Benchmark. – Cité pages 27, 28, 29, 30, 31, 34, 37, 45, 46, 66, 71, 107, 108, 113 et 121.

Han, X., Chen, K., Zhou, Y., Qiu, M., Fan, C., Liu, Y., and Zhang, T. (2021b). A Unified Anomaly Detection Methodology for Lane-Following of Autonomous Driving Systems. In *2021 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*, pages 836–844. – Cité page 54.

Harman, D., Ahmed, M., Mahmood, A. N., and Islam, M. R. (2002). Overview of the TREC 2002 novelty track. In *Proceedings of the Eleventh Text REtrieval Conference (TREC 2002), NIST Special Publication 500-251*. Citeseer. – Cité pages 64 et 109.

Harmouche, J., Delpha, C., and Diallo, D. (2014). Incipient fault detection and diagnosis based on Kullback–Leibler divergence using Principal Component Analysis: Part I. *Signal Processing*, 94:278–287. – Cité page 44.

Harmouche, J., Delpha, C., and Diallo, D. (2015). Incipient fault detection and diagnosis based on Kullback–Leibler divergence using principal component analysis: Part II. *Signal Processing*, 109:334–344. – Cité page 44.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108. – Cité pages 35 et 40.

Hassan, A. K. I. and Abraham, A. (2016). Modeling Insurance Fraud Detection Using Imbalanced Data Classification. In Pillay, N., Engelbrecht, A. P., Abraham, A., du Plessis, M. C., Snášel, V., and Muda, A. K., editors, *Advances in Nature and Biologically Inspired Computing*, Advances in Intelligent Systems and Computing, pages 117–127, Cham. Springer International Publishing. – Cité page 28.

Hauskrecht, M., Batal, I., Valko, M., Visweswaran, S., Cooper, G. F., and Clermont, G. (2013). Outlier detection for patient monitoring and alerting. *Journal of Biomedical Informatics*, 46(1):47–55. – Cité page 20.

Hawkins, D. M. (1974). The Detection of Errors in Multivariate Data Using Principal Components. *Journal of the American Statistical Association*, 69(346):340–344. – Cité page 26.

Hawkins, D. M. (1980). *Identification of Outliers*. Springer Netherlands, Dordrecht. – Cité pages 21 et 22.

He, H., Wang, J., Graco, W., and Hawkins, S. (1997). Application of neural networks to detection of medical fraud. *Expert Systems with Applications*, 13(4):329–336. – Cité page 26.

He, Z., Xu, X., and Deng, S. (2003). Discovering cluster-based local outliers. *Pattern Recognition Letters*, 24(9-10):1641–1650. – Cité pages 35 et 40.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. (2021). Measuring massive multitask language understanding. In *International Conference on Learning Representations*. – Cité page 187.

Hilal, W., Gadsden, S. A., and Yawney, J. (2022). Financial Fraud: A Review of Anomaly Detection Techniques and Recent Advances. *Expert Systems with Applications*, 193:116429. – Cité page 20.

Hiltunen, E. (2007). Where do Future-Oriented People Find Weak. *FFRC eBook*, page 64. – Cité page 73.

Hoang, D. H. and Nguyen, H. D. (2018). A PCA-based method for IoT network traffic anomaly detection. In *2018 20th International Conference on Advanced Communication Technology (ICACT)*, pages 381–386. – Cité pages 35 et 42.

Hoffmann, H. (2007). Kernel PCA for novelty detection. *Pattern recognition*, 40(3):863–874. – Cité pages 21, 35 et 42.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J. W., Vinyals, O., and Sifre, L. (2022). Training Compute-Optimal Large Language Models. – Cité page 164.

Hollier, G. and Austin, J. (2002). Novelty detection for strain-gauge degradation using maximally correlated components. In *ESANN'2002 Proceedings*, pages 257–262. – Cité page 38.

Homayouni, H., Ray, I., Ghosh, S., Gondalia, S., and Kahn, M. G. (2021). Anomaly Detection in COVID-19 Time-Series Data. *SN Computer Science*, 2(4):279. – Cité page 52.

Hong, D., Zhao, D., and Zhang, Y. (2016). The Entropy and PCA Based Anomaly Prediction in Data Streams. *Procedia Computer Science*, 96:139–146. – Cité page 44.

Howedi, A., Lotfi, A., and Pourabdollah, A. (2020). An entropy-based approach for anomaly detection in activities of daily living in the presence of a visitor. *Entropy. An International and Interdisciplinary Journal of Entropy and Information Studies*, 22(8):845. – Cité page 44.

Hu, C., Feng, Y., Kamigaito, H., Takamura, H., and Okumura, M. (2021). One-class Text Classification with Multi-modal Deep Support Vector Data Description. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, volume Main Volume, pages 3378–3390. Association for Computational Linguistics. – Cité page 66.

Hu, C., Wu, J., Sun, C., Chen, X., and Yan, R. (2023). Mutual information-based feature disentangled network for anomaly detection under variable working conditions. *Mechanical Systems and Signal Processing*, 204:110804. – Cité page 44.

Huang, C., Ye, F., Zhao, P., Zhang, Y., Wang, Y.-F., and Tian, Q. (2021). ESAD: End-to-end Deep Semi-supervised Anomaly Detection. – Cité page 44.

Huang, D., Mu, D., Yang, L., and Cai, X. (2018). CoDetect: Financial Fraud Detection With Anomaly Feature Detection. *IEEE Access*, 6:19161–19174. – Cité page 20.

Huang, H., Zhang, B., Sun, Y., Ma, C., and Qu, J. (2022). Delta-DAGMM: A Free Rider Attack Detection Model in Horizontal Federated Learning. *Security and Communication Networks*, 2022(1):8928790. – Cité page 39.

Huang, T., Zhu, Y., Zhang, Q., Zhu, Y., Wang, D., Qiu, M., and Liu, L. (2013). An LOF-Based Adaptive Anomaly Detection Scheme for Cloud Computing. In *2013 IEEE 37th Annual Computer Software and Applications Conference Workshops*, pages 206–211. – Cité page 40.

Hundman, K., Constantinou, V., Laporte, C., Colwell, I., and Soderstrom, T. (2018). Detecting Spacecraft Anomalies Using LSTMs and Nonparametric Dynamic Thresholding. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pages 387–395, New York, NY, USA. Association for Computing Machinery. – Cité page 26.

Hutwagner, L., Browne, T., Seeman, G. M., and Fleischauer, A. T. (2005). Comparing aberration detection methods with simulated data. *Emerging infectious diseases*, 11(2):314. – Cité page 21.

Iacus, S. M., Sermi, F., Spyratos, S., Tarchi, D., and Vespe, M. (2021). Anomaly detection of mobile positioning data with applications to COVID-19 situational awareness. *Japanese Journal of Statistics and Data Science*, 4(1):763–781. – Cité page 53.

Ikram, S. T., Cherukuri, A. K., Poorva, B., Ushasree, P. S., Zhang, Y., Liu, X., and Li, G. (2021). Anomaly Detection Using XGBoost Ensemble of Deep Neural Network Models. *Cybernetics and Information Technologies*, 21(3):175–188. – Cité page 45.

Ilgun, K., Kemmerer, R., and Porras, P. (1995). State transition analysis: A rule-based intrusion detection approach. *IEEE Transactions on Software Engineering*, 21(3):181–199. – Cité page 26.

Iqbal, A. and Amin, R. (2024). Time series forecasting and anomaly detection using deep learning. *Computers & Chemical Engineering*, 182:108560. – Cité page 52.

Islam, R., Refat, R. U. D., Yerram, S. M., and Malik, H. (2022). Graph-Based Intrusion Detection System for Controller Area Networks. *IEEE Transactions on Intelligent Transportation Systems*, 23(3):1727–1736. – Cité page 54.

Itti, L. and Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research*, 49(10):1295–1306. – Cité page 21.

Jackson, M. L., Baer, A., Painter, I., and Duchin, J. (2007). A simulation study comparing aberration detection algorithms for syndromic surveillance. *BMC medical informatics and decision making*, 7:1–11. – Cité page 21.

Jafari, A. (2022). A Deep Learning Anomaly Detection Method in Textual Data. – Cité page 65.

Jain, P. K., Bajpai, M. S., and Pamula, R. (2022). A modified DBSCAN algorithm for anomaly detection in time-series data with seasonality. *Int. Arab J. Inf. Technol.*, 19(1):23–28. – Cité page 40.

Jain, R. B. (2010). A recursive version of Grubbs' test for detecting multiple outliers in environmental and chemical data. *Clinical Biochemistry*, 43(12):1030–1033. – Cité page 37.

Janetzko, H., Stoffel, F., Mittelstädt, S., and Keim, D. A. (2014). Anomaly detection for visual analytics of power consumption data. *Computers & Graphics*, 38:27–37. – Cité page 52.

Janjua, Z. H., Vecchio, M., Antonini, M., and Antonelli, F. (2019). IRESE: An intelligent rare-event detection system using unsupervised learning on the IoT edge. *Engineering Applications of Artificial Intelligence*, 84:41–50. – Cité pages 20 et 21.

Jaskie, K. and Spanias, A. (2022). *Positive Unlabeled Learning*. Morgan & Claypool Publishers. – Cité page 32.

Jiang, F., Sui, Y., and Cao, C. (2010). An information entropy-based approach to outlier detection in rough sets. *Expert Systems with Applications*, 37(9):6338–6344. – Cité page 44.

Jiang, M., Hou, C., Zheng, A., Hu, X., Han, S., Huang, H., He, X., Yu, P. S., and Zhao, Y. (2023). Weakly Supervised Anomaly Detection: A Survey. https://arxiv.org/abs/2302.04549v1. – Cité pages 29, 31 et 46.

Jiang, W., Hong, Y., Zhou, B., He, X., and Cheng, C. (2019). A GAN-Based Anomaly Detection Approach for Imbalanced Industrial Time Series. *IEEE Access*, 7:143608–143619. – Cité page 43.

Jombart, T., Ghozzi, S., Schumacher, D., Taylor, T. J., Leclerc, Q. J., Jit, M., Flasche, S., Greaves, F., Ward, T., Eggo, R. M., Nightingale, E., Meakin, S., Brady, O. J., null, n., Medley, G. F., Höhle, M., and Edmunds, W. J. (2021). Real-time monitoring of COVID-19 dynamics using automated trend fitting and anomaly detection. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 376(1829):20200266. – Cité page 52.

Jordan, M. I. and Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260. – Cité page 28.

Kale, R. and Thing, V. L. L. (2023). Few-shot weakly-supervised cybersecurity anomaly detection. *Computers & Security*, 130:103194. – Cité page 34.

Kamaruddin, S. S., Bakar, A. A., Hamdan, A. R., Nor, F. M., Nazri, M. Z. A., Othman, Z. A., and Hussein, G. S. (2015). A text mining system for deviation detection in financial documents. *Intelligent Data Analysis*, 19(s1):S19–S44. – Cité page 21.

Kanda, Y., Fontugne, R., Fukuda, K., and Sugawara, T. (2013). ADMIRE: Anomaly detection method using entropy-based PCA with three-step sketches. *Computer Communications*, 36(5):575–588. – Cité page 44.

Kannan, R., Woo, H., Aggarwal, C. C., and Park, H. (2017). Outlier Detection for Text Data : An Extended Version. *arXiv:1701.01325 [cs, stat]*. – Cité page 21.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling Laws for Neural Language Models. – Cité pages 163 et 164.

Kapusi, T. P., Kovács, L., and Hajdu, A. (2022). Deep learning-based anomaly detection for imaging in autonomous vehicles. In *2022 IEEE 2nd Conference on Information Technology and Data Science (CITDS)*, pages 142–147. – Cité page 54.

Karadayi, Y., Aydin, M. N., and Öğrencí, A. S. (2020). Unsupervised Anomaly Detection in Multivariate Spatio-Temporal Data Using Deep Learning: Early Detection of COVID-19 Outbreak in Italy. *IEEE Access*, 8:164155–164177. – Cité page 53.

Kawachi, Y., Koizumi, Y., and Harada, N. (2018). Complementary set variational autoencoder for supervised anomaly detection. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2366–2370. IEEE. – Cité page 30.

Keung, P., Lu, Y., Szarvas, G., and Smith, N. A. (2020). The Multilingual Amazon Reviews Corpus. *arXiv:2010.02573 [cs]*. – Cité pages 66 et 110.

Khaleghi, A. and Moin, M. S. (2018). Improved anomaly detection in surveillance videos based on a deep learning method. In *2018 8th Conference of AI & Robotics and 10th RoboCup Iranopen International Symposium (IRANOPEN)*, pages 73–81. – Cité page 54.

Khan, S. S. and Madden, M. G. (2014). One-class classification: Taxonomy of study and review of techniques. *The Knowledge Engineering Review*, 29(3):345–374. – Cité page 31.

Khan, S. W., Hafeez, Q., Khalid, M. I., Alroobaea, R., Hussain, S., Iqbal, J., Almotiri, J., and Ullah, S. S. (2022). Anomaly Detection in Traffic Surveillance Videos Using Deep Learning. *Sensors*, 22(17):6563. – Cité page 54.

Khan, W. and Haroon, M. (2022). An unsupervised deep learning ensemble model for anomaly detection in static attributed social networks. *International Journal of Cognitive Computing in Engineering*, 3:153–160. – Cité page 36.

Kieu, T., Yang, B., Guo, C., and Jensen, C. S. (2019). Outlier Detection for Time Series with Recurrent Autoencoder Ensembles. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 2725–2732, Macao, China. International Joint Conferences on Artificial Intelligence Organization. – Cité page 42.

Kim, D., Cha, J., Oh, S., and Jeong, J. (2021a). AnoGAN-Based Anomaly Filtering for Intelligent Edge Device in Smart Factory. In *2021 15th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, pages 1–6. – Cité page 43.

Kim, J., Ko, J., Choi, H., and Kim, H. (2021b). Printed Circuit Board Defect Detection Using Deep Learning via A Skip-Connected Convolutional Autoencoder. *Sensors*, 21(15):4968. – Cité page 53.

Kim, J. and Lee, C. (2017). Novelty-focused weak signal detection in futuristic data: Assessing the rarity and paradigm unrelatedness of signals. *Technological Forecasting and Social Change*, 120:59–76. – Cité page 20.

Kind, A., Stoecklin, M. P., and Dimitropoulos, X. (2009). Histogram-based traffic anomaly detection. *IEEE Transactions on Network and Service Management*, 6(2):110–121. – Cité page 38.

Kiran, B., Thomas, D., and Parakkal, R. (2018). An Overview of Deep Learning Based Methods for Unsupervised and Semi-Supervised Anomaly Detection in Videos. *Journal of Imaging*, 4(2):36. – Cité page 31.

Knapp, O., Cerri, O., Dissertori, G., Nguyen, T. Q., Pierini, M., and Vlimant, J. R. (2021). Adversarially Learned Anomaly Detection on CMS open data: Rediscovering the top quark. *The European Physical Journal Plus*, 136(2):236. – Cité page 43.

Knorr, E. M. (2002). *Outliers and Data Mining: Finding Exceptions in Data*. PhD thesis, University of British Columbia. – Cité page 21.

Kong, X., Gao, H., Alfarraj, O., Ni, Q., Zheng, C., and Shen, G. (2020). HUAD: Hierarchical Urban Anomaly Detection Based on Spatio-Temporal Data. *IEEE Access*, 8:26573–26582. – Cité page 53.

Kopylova, Y., Buell, D. A., Huang, C.-T., and Janies, J. (2008). Mutual information applied to anomaly detection. *Journal of Communications and Networks*, 10(1):89–97. – Cité page 44.

Kramer, S. (2010). Anomaly detection in extremist web forums using a dynamical systems approach. In *ACM SIGKDD Workshop on Intelligence and Security Informatics*, pages 1–10, Washington D.C. ACM. – Cité page 60.

Kriegel, H.-P., Kröger, P., Schubert, E., and Zimek, A. (2009). LoOP: Local outlier probabilities. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 1649–1652, New York, NY, USA. Association for Computing Machinery. – Cité page 40.

Kriegel, H.-P., Schubert, M., and Zimek, A. (2008). Angle-based outlier detection in high-dimensional data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 444–452, Las Vegas Nevada USA. ACM. – Cité pages 40, 86 et 103.

Kumar, B. B. S. and Ravi, V. (2017). Text Document Classification with PCA and One-Class SVM. In *Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications*, volume 515, pages 107–115. Springer Singapore, Singapore. – Cité page 42.

Kumar, R., Jain, A., Tripathi, A. K., and Tyagi, S. (2021). COVID-19 Outbreak: An Epidemic Analysis using Time Series Prediction Model. In *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pages 1090–1094. – Cité page 52.

Kumar, S., Khan, M. B., Hasanat, M. H. A., Saudagar, A. K. J., AlTameem, A., and AlKhathami, M. (2022). An Anomaly Detection Framework for Twitter Data. *Applied Sciences*, 12(21):11059. – Cité page 69.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2020). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *arXiv:1909.11942 [cs]*. – Cité page 108.

Landauer, M., Onder, S., Skopik, F., and Wurzenberger, M. (2023). Deep Learning for Anomaly Detection in Log Data: A Survey. *Machine Learning with Applications*, 12:100470. – Cité page 37.

Lang, C. I., Sun, F.-K., Lawler, B., Dillon, J., Dujaili, A. A., Ruth, J., Cardillo, P., Alfred, P., Bowers, A., Mckiernan, A., and Boning, D. S. (2022). One Class Process Anomaly Detection Using Kernel Density Estimation Methods. *IEEE Transactions on Semiconductor Manufacturing*, 35(3):457–469. – Cité page 39.

Laorden, C., Ugarte-Pedrero, X., Santos, I., Sanz, B., Nieves, J., and Bringas, P. G. (2014). Study on the effectiveness of anomaly detection for spam filtering. *Information Sciences*, 277:421–444. – Cité page 60.

Latecki, L. J., Lazarevic, A., and Pokrajac, D. (2007). Outlier Detection with Kernel Density Functions. In Perner, P., editor, *Machine Learning and Data Mining in Pattern Recognition*, volume 4571, pages 61–75. Springer Berlin Heidelberg, Berlin, Heidelberg. – Cité page 38.

Laxhammar, R., Falkman, G., and Sviestins, E. (2009). Anomaly detection in sea traffic - A comparison of the Gaussian Mixture Model and the Kernel Density Estimator. In *2009 12th International Conference on Information Fusion*, pages 756–763. – Cité page 38.

Le, Q. V. and Mikolov, T. (2014). Distributed Representations of Sentences and Documents. – Cité pages 68 et 83.

Lei, Z., Zhu, L., Fang, Y., Li, X., and Liu, B. (2020). Anomaly detection of bridge health monitoring data based on KNN algorithm. *Journal of Intelligent & Fuzzy Systems*, 39(4):5243–5252. – Cité page 39.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics. – Cité page 159.

Li, D., Guo, H., Wang, Z., and Zheng, Z. (2021). Unsupervised Fake News Detection Based on Autoencoder. *IEEE Access*, 9:29356–29365. – Cité page 60.

Li, G. and Jung, J. J. (2021). Entropy-based dynamic graph embedding for anomaly detection on multiple climate time series. *Scientific Reports*, 11(1):13819. – Cité page 52.

Li, H. (2024). *Machine Learning Methods*. Springer Nature, Singapore. – Cité page 28.

Li, H. and Li, Y. (2023). Anomaly detection methods based on GAN: A survey. *Applied Intelligence*, 53(7):8209–8231. – Cité page 43.

Li, K., Ling, Q., Qin, Y., Wang, Y., Cai, Y., Lin, Z., and An, W. (2022a). Spectral-Spatial Deep Support Vector Data Description for Hyperspectral Anomaly Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16. – Cité page 41.

Li, K.-L., Huang, H.-K., Tian, S.-F., and Xu, W. (2003). Improving one-class SVM for anomaly detection. In *Proceedings of the 2003 International Conference on Machine Learning and Cybernetics (IEEE Cat. No.03EX693)*, volume 5, pages 3077–3081 Vol.5. – Cité page 41.

Li, L., Hansman, R. J., Palacios, R., and Welsch, R. (2016). Anomaly detection via a Gaussian Mixture Model for flight operation and safety monitoring. *Transportation Research Part C: Emerging Technologies*, 64:45–57. – Cité page 39.

Li, R., Liu, Z., Ma, Y., Yang, D., and Sun, S. (2023). Internet Financial Fraud Detection Based on Graph Learning. *IEEE Transactions on Computational Social Systems*, 10(3):1394–1401. – Cité page 54.

Li, T., Comer, M. L., Delp, E. J., Desai, S. R., Mathieson, J. L., Foster, R. H., and Chan, M. W. (2020a). Anomaly Scoring for Prediction-Based Anomaly Detection in Time Series. In *2020 IEEE Aerospace Conference*, pages 1–7. – Cité page 26.

Li, Z., Li, Y., and Xu, L. (2011). Anomaly intrusion detection method based on k-means clustering algorithm with particle swarm optimization. In *2011 International Conference of Information Technology, Computer Engineering and Management Sciences*, volume 2, pages 157–161. IEEE. – Cité page 35.

Li, Z., Sun, C., Liu, C., Chen, X., Wang, M., and Liu, Y. (2022b). Dual-MGAN: An Efficient Approach for Semi-supervised Outlier Detection with Few Identified Anomalies. *ACM Trans. Knowl. Discov. Data*, 16(6):107:1–107:30. – Cité page 34.

Li, Z., Zhao, Y., Botta, N., Ionescu, C., and Hu, X. (2020b). COPOD: Copula-Based Outlier Detection. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 1118–1123, Sorrento, Italy. IEEE. – Cité pages 35, 38, 88 et 103.

Li, Z., Zhao, Y., Hu, X., Botta, N., Ionescu, C., and Chen, G. H. (2022c). ECOD: Unsupervised Outlier Detection Using Empirical Cumulative Distribution Functions. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1. – Cité pages 35, 38, 90 et 103.

Likas, A., Vlassis, N., and J. Verbeek, J. (2003). The global k-means clustering algorithm. *Pattern Recognition*, 36(2):451–461. – Cité page 40.

Lim, S. K., Loo, Y., Tran, N.-T., Cheung, N.-M., Roig, G., and Elovici, Y. (2018). DOPING: Generative Data Augmentation for Unsupervised Anomaly Detection with GAN. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 1122–1127. – Cité page 64.

Lin, J., Keogh, E., Fu, A., and Van Herle, H. (2005). Approximations to magic: Finding unusual medical time series. In *18th IEEE Symposium on Computer-Based Medical Systems (CBMS'05)*, pages 329–334. – Cité page 26.

Lin, X., Wang, H., Guo, J., and Mei, G. (2022). A Deep Learning Approach Using Graph Neural Networks for Anomaly Detection in Air Quality Data Considering Spatiotemporal Correlations. *IEEE Access*, 10:94074–94088. – Cité page 53.

Liu, B., Li, X., Xiao, Y., Sun, P., Zhao, S., Peng, T., Zheng, Z., and Huang, Y. (2024). Adaboost-based SVDD for anomaly detection with dictionary learning. *Expert Systems with Applications*, 238:121770. – Cité page 41.

Liu, C. and Gryllias, K. (2020). A semi-supervised Support Vector Data Description-based fault detection method for rolling element bearings based on cyclic spectral analysis. *Mechanical Systems and Signal Processing*, 140:106682. – Cité pages 31 et 41.

Liu, F., Yu, Y., Song, P., Fan, Y., and Tong, X. (2020a). Scalable KDE-based top-n local outlier detection over large-scale data streams. *Knowledge-Based Systems*, 204:106186. – Cité pages 38 et 39.

Liu, F. T., Ting, K. M., and Zhou, Z.-H. (2008). Isolation Forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422, Pisa, Italy. IEEE. – Cité pages 45 et 122.

Liu, F. T., Ting, K. M., and Zhou, Z.-H. (2012). Isolation-Based Anomaly Detection. *ACM Transactions on Knowledge Discovery from Data*, 6(1):1–39. – Cité page 45.

Liu, G., Niu, Y., Zhao, W., Duan, Y., and Shu, J. (2022a). Data anomaly detection for structural health monitoring using a combination network of GANomaly and CNN. *Smart Structures and Systems*, 29(1):53–62. – Cité page 43.

Liu, J., Zhu, H., Liu, Y., Wu, H., Lan, Y., and Zhang, X. (2019a). Anomaly detection for time series using temporal convolutional networks and Gaussian mixture model. *Journal of Physics: Conference Series*, 1187(4):042111. – Cité page 39.

Liu, L., Wang, P., Lin, J., and Liu, L. (2021a). Intrusion Detection of Imbalanced Network Traffic Based on Machine Learning and Deep Learning. *IEEE Access*, 9:7550–7563. – Cité page 28.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2021b). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *arXiv:2107.13586 [cs]*. – Cité pages 70, 158, 159 et 165.

Liu, W., Jiang, H., Che, D., Chen, L., and Jiang, Q. (2020b). A Real-time Temperature Anomaly Detection Method for IoT Data. In *Proceedings of the 5th International Conference on Internet of Things, Big Data and Security*, pages 112–118, Prague, Czech Republic. SCITEPRESS - Science and Technology Publications. – Cité page 20.

Liu, X., Ding, Y., Tang, H., and Xiao, F. (2021c). A data mining-based framework for the identification of daily electricity usage patterns and anomaly detection in building electricity consumption data. *Energy and Buildings*, 231:110601. – Cité page 52.

Liu, Y., He, R., Qu, Y., Zhu, Y., Li, D., Ling, X., Xia, S., Li, Z., and Li, D. (2022b). Integration of Human Protein Sequence and Protein-Protein Interaction Data by Graph Autoencoder to Identify Novel Protein-Abnormal Phenotype Associations. *Cells*, 11(16):2485. – Cité page 54.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019b). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*. – Cité page 70.

Lokanan, M., Tran, V., and Vuong, N. H. (2019). Detecting anomalies in financial statements using machine learning algorithm: The case of Vietnamese listed firms. *Asian Journal of Accounting Research*, 4(2):181–201. – Cité page 60.

Luo, C., Zhao, Y., Cao, L., Ou, Y., and Zhang, C. (2008). Exception mining on multiple time series in stock market. In *2008 IEEE / WIC / ACM International Conference on Web Intelligence and Intelligent Agent Technology*, volume 3, pages 690–693. IEEE. – Cité pages 20, 21 et 52.

Luo, Z., Zuo, R., Xiong, Y., and Wang, X. (2021). Detection of geochemical anomalies related to mineralization using the GANomaly network. *Applied Geochemistry*, 131:105043. – Cité pages 32 et 43.

Lyu, J. and Manoochehri, S. (2021). Online Convolutional Neural Network-based anomaly detection and quality control for Fused Filament Fabrication process. *Virtual and Physical Prototyping*, 16(2):160–177. – Cité page 53.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, page 9. – Cité pages 66 et 110.

Madan, V., Khetan, A., and Karnin, Z. (2021). TADPOLE: Task ADapted Pre-Training via AnOmaLy DEtection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5732–5746, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. – Cité pages 66 et 69.

Mai, K. T., Davies, T., and Griffin, L. D. (2022). Self-Supervised Losses for One-Class Textual Anomaly Detection. – Cité pages 66 et 71.

Makridis, G., Kyriazis, D., and Plitsos, S. (2020). Predictive maintenance leveraging machine learning for time-series forecasting in the maritime industry. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–8. – Cité page 52.

Manolache, A., Brad, F., and Burceanu, E. (2021). DATE: Detecting Anomalies in Text via Self-Supervision of Transformers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 267–277, Online. Association for Computational Linguistics. – Cité pages 31, 66, 71, 106 et 113.

Markou, M. and Singh, S. (2003a). Novelty detection: A review—part 1: Statistical approaches. *Signal Processing*, 83(12):2481–2497. – Cité pages 21, 31, 35 et 37.

Markou, M. and Singh, S. (2003b). Novelty detection: A review—part 2: Neural network based approaches. *Signal Processing*, 83(12):2499–2521. – Cité pages 21, 31, 34 et 37.

Marzuoli, A. and Liu, F. (2019). Monitoring of natural disasters through anomaly detection on mobile phone data. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 4089–4098. – Cité page 53.

Massoli, F. V., Falchi, F., Kantarci, A., Akti, Ş., Ekenel, H. K., and Amato, G. (2022). MOCCA: Multilayer One-Class Classification for Anomaly Detection. *IEEE Transactions on Neural Networks and Learning Systems*, 33(6):2313–2323. – Cité page 32.

Matoušek, J. and Tihelka, D. (2017). Annotation Error Detection: Anomaly Detection vs. Classification. In Karpov, A., Potapova, R., and Mporas, I., editors, *Speech and Computer*, pages 141–151, Cham. Springer International Publishing. – Cité page 60.

Mei, M., Guo, X., Williams, B. C., Doboli, S., Kenworthy, J. B., Paulus, P. B., and Minai, A. A. (2018). Using Semantic Clustering And Autoencoders For Detecting Novelty In Corpora Of Short Texts. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, Rio de Janeiro. IEEE. – Cité page 67.

Meng, W., Liu, Y., Zhu, Y., Zhang, S., Pei, D., Liu, Y., Chen, Y., Zhang, R., Tao, S., Sun, P., and Zhou, R. (2019). LogAnomaly: Unsupervised Detection of Sequential and Quantitative Anomalies in Unstructured Logs. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 4739–4745, Macao, China. International Joint Conferences on Artificial Intelligence Organization. – Cité page 20.

Miao, X., Liu, Y., Zhao, H., and Li, C. (2019). Distributed Online One-Class Support Vector Machine for Anomaly Detection Over Networks. *IEEE Transactions on Cybernetics*, 49(4):1475–1488. – Cité page 41.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. – Cité pages 68, 82 et 157.

Miljković, D. (2010). Review of Novelty Detection Methods. page 6. – Cité pages 35 et 37.

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., and Zettlemoyer, L. (2022a). Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? – Cité page 176.

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., and Zettlemoyer, L. (2022b). Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? – Cité page 192.

Moschini, G., Houssou, R., Bovay, J., and Robert-Nicoud, S. (2021). Anomaly and Fraud Detection in Credit Card Transactions Using the ARIMA Model. *Engineering Proceedings*, 5(1):56. – Cité pages 20 et 52.

Moso, J. C., Cormier, S., de Runz, C., Fouchal, H., and Wandeto, J. M. (2021). Anomaly Detection on Data Streams for Smart Agriculture. *Agriculture*, 11(11):1083. – Cité page 53.

Mu, H., Sun, R., Yuan, G., and Shi, G. (2021a). Positive unlabeled learning-based anomaly detection in videos. *International Journal of Intelligent Systems*, 36(8):3767–3788. – Cité page 32.

Mu, J., Zhang, X., Li, Y., and Guo, J. (2021b). Deep neural network for text anomaly detection in SIoT. *Computer Communications*, 178:286–296. – Cité page 42.

Mukherjee, P., Roy, C. K., and Roy, S. K. (2022). Ocformer: One-class transformer network for image classification. *arXiv preprint arXiv:2204.11449*. – Cité page 31.

Munz, G., Li, S., and Carle, G. (2007). Traffic Anomaly Detection Using K-Means Clustering. In *GI/ITG Workshop MMBnet*, page 8. – Cité pages 35 et 40.

Müter, M. and Asaj, N. (2011). Entropy-based anomaly detection for in-vehicle networks. In *2011 IEEE Intelligent Vehicles Symposium (IV)*, pages 1110–1115. – Cité page 44.

Na, G. S., Kim, D., and Yu, H. (2018). Dilof: Effective and memory efficient local outlier detection in data streams. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1993–2002. – Cité pages 35 et 40.

Napoletano, P., Piccoli, F., and Schettini, R. (2021). Semi-supervised anomaly detection for visual quality inspection. *Expert Systems with Applications*, 183:115275. – Cité page 53.

Nassif, A. B., Talib, M. A., Nasir, Q., and Dakalbab, F. M. (2021). Machine Learning for Anomaly Detection: A Systematic Review. *IEEE Access*, 9:78658–78700. – Cité pages 27, 36, 37 et 42.

Navaz, A. S. S., Sangeetha, V., and Prabhadevi, C. (2013). Entropy based Anomaly Detection System to Prevent DDoS Attacks in Cloud. – Cité page 44.

Nayak, R., Behera, M. M., Pati, U. C., and Das, S. K. (2019). Video-based Real-time Intrusion Detection System using Deep-Learning for Smart City Applications. In *2019 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS)*, pages 1–6. – Cité page 54.

Nguyen, K.-T., Dinh, D.-T., Do, M. N., and Tran, M.-T. (2020). Anomaly Detection in Traffic Surveillance Videos with GAN-based Future Frame Prediction. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, ICMR '20, pages 457–463, New York, NY, USA. Association for Computing Machinery. – Cité page 54.

Nkenyereye, L., Tama, B., and Lim, S. (2020). A Stacking-Based Deep Neural Network Approach for Effective Network Anomaly Detection. *Computers, Materials & Continua*, 66(2):2217–2227. – Cité page 45.

Nun, I., Protopapas, P., Sim, B., and Chen, W. (2016). Ensemble Learning Method for Outlier Detection and its Application to Astronomical Light Curves. *The Astronomical Journal*, 152(3):71. – Cité page 45.

Omar, S., Ngadi, A., and H. Jebur, H. (2013). Machine Learning Techniques for Anomaly Detection: An Overview. *International Journal of Computer Applications*, 79(2):33–41. – Cité pages 27 et 28.

Ortega Vázquez, C., vanden Broucke, S., and De Weerdt, J. (2023). A two-step anomaly detection based method for PU classification in imbalanced data sets. *Data Mining and Knowledge Discovery*, 37(3):1301–1325. – Cité page 32.

Ousidhoum, N., Lin, Z., Zhang, H., Song, Y., and Yeung, D.-Y. (2019). Multilingual and Multi-Aspect Hate Speech Analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4674–4683, Hong Kong, China. Association for Computational Linguistics. – Cité page 112.

Ouyang, B., Song, Y., Li, Y., Sant, G., and Bauchy, M. (2021). EBOD: An ensemble-based outlier detection algorithm for noisy datasets. *Knowledge-Based Systems*, 231:107400. – Cité page 45.

Ouyang, Z., Sun, X., Chen, J., Yue, D., and Zhang, T. (2018). Multi-view stacking ensemble for power consumption anomaly detection in the context of industrial internet of things. *IEEE access : practical innovations, open solutions*, 6:9623–9631. – Cité page 45.

Oza, P. and Patel, V. M. (2019). One-Class Convolutional Neural Network. *IEEE Signal Processing Letters*, 26(2):277–281. – Cité page 32.

Palpanas, T., Papadopoulos, D., Kalogeraki, V., and Gunopulos, D. (2003). Distributed deviation detection in sensor networks. *Acm sigmod record*, 32(4):77–82. – Cité page 21.

Pan, H., Badawi, D., Bassi, I., Ozev, S., and Cetin, A. E. (2022). Detecting Anomaly in Chemical Sensors via L1-Kernel-Based Principal Component Analysis. *IEEE Sensors Letters*, 6(10):1–4. – Cité page 42.

Pang, G., Cao, L., and Aggarwal, C. (2021a). Deep Learning for Anomaly Detection: Challenges, Methods, and Opportunities. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 1127–1130, Virtual Event Israel. ACM. – Cité page 27.

Pang, G., Cao, L., Chen, L., and Liu, H. (2018). Learning Representations of Ultrahigh-dimensional Data for Random Distance-based Outlier Detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2041–2050. – Cité page 34.

Pang, G., Shen, C., Cao, L., and van den Hengel, A. (2021b). Deep Learning for Anomaly Detection: A Review. *ACM Computing Surveys*, 54(2):1–38. – Cité pages 21, 27, 30, 37 et 46.

Pang, G., Shen, C., Jin, H., and van den Hengel, A. (2020). Deep Weakly-supervised Anomaly Detection. – Cité pages 29, 31, 46 et 103.

Pang, G., Shen, C., Jin, H., and van den Hengel, A. (2023). Deep Weakly-supervised Anomaly Detection. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, pages 1795–1807, New York, NY, USA. Association for Computing Machinery. – Cité pages 31, 36 et 99.

Pang, G., Shen, C., and Van Den Hengel, A. (2019). Deep anomaly detection with deviation networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 353–362. – Cité pages 36, 46, 64, 95, 99 et 103.

Pantin, J., Lesot, M.-J., and Marsala, C. (2022). Analyse de données aberrantes pour le texte: Taxonomie et étude expérimentale. *TextMine'22*. – Cité pages 65, 107 et 113.

Papadimitriou, S., Kitagawa, H., Gibbons, P., and Faloutsos, C. (2003). LOCI: Fast outlier detection using the local correlation integral. In *Proceedings 19th International Conference on Data Engineering (Cat. No.03CH37405)*, pages 315–326, Bangalore, India. IEEE. – Cité pages 35 et 40.

Pascoal, C., de Oliveira, M. R., Valadas, R., Filzmoser, P., Salvador, P., and Pacheco, A. (2012). Robust feature selection and robust PCA for internet traffic anomaly detection. In *2012 Proceedings IEEE INFOCOM*, pages 1755–1763, Orlando, FL, USA. IEEE. – Cité pages 20, 35 et 42.

Pasillas-Díaz, J. R. and Ratté, S. (2017). Bagged Subspaces for Unsupervised Outlier Detection. *Computational Intelligence*, 33(3):507–523. – Cité page 45.

Patcha, A. and Park, J.-M. (2007). An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Computer Networks*, 51(12):3448–3470. – Cité page 37.

Paulauskas, N. and Baskys, A. (2019). Application of Histogram-Based Outlier Scores to Detect Computer Network Anomalies. *Electronics*, 8(11):1251. – Cité page 38.

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. – Cité page 113.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics. – Cité pages 68, 82 et 157.

Perera, P., Oza, P., and Patel, V. M. (2021). One-Class Classification: A Survey. *arXiv:2101.03064 [cs]*. – Cité page 31.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics. – Cité page 82.

Pevný, T. (2016). Loda: Lightweight on-line detector of anomalies. *Machine Learning*, 102(2):275–304. – Cité pages 35 et 40.

Pham, L., Phan, H., Palaniappan, R., Mertins, A., and McLoughlin, I. (2021). CNN-MoE Based Framework for Classification of Respiratory Anomalies and Lung Disease Detection. *IEEE Journal of Biomedical and Health Informatics*, 25(8):2938–2947. – Cité page 45.

Pimentel, M. A., Clifton, D. A., Clifton, L., and Tarassenko, L. (2014). A review of novelty detection. *Signal Processing*, 99:215–249. – Cité pages 21, 30, 31, 32, 35, 37, 39, 41, 42 et 43.

Presbitero, A., Quax, R., Krzhizhanovskaya, V., and Sloot, P. (2017). Anomaly Detection in Clinical Data of Patients Undergoing Heart Surgery. *Procedia Computer Science*, 108:99–108. – Cité page 52.

Qian, J., Li, X., Liao, S., and Yeh, A.-G.-o. (2010). Applying an Anomaly-Detection Algorithm for Short-Term Land Use and Land Cover Change Detection Using Time-Series SAR Images. *GIScience & Remote Sensing*, 47(3):379–397. – Cité page 53.

Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., and Huang, X. (2020). Pre-trained Models for Natural Language Processing: A Survey. *Science China Technological Sciences*, 63(10):1872–1897. – Cité page 165.

Qu, J., Du, Q., Li, Y., Tian, L., and Xia, H. (2021). Anomaly Detection in Hyperspectral Imagery Based on Gaussian Mixture Model. *IEEE Transactions on Geoscience and Remote Sensing*, 59(11):9504–9517. – Cité page 38.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training. page 12. – Cité page 168.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. *OpenAI blog*, 1:24. – Cité pages 159, 164 et 169.

Ramaswamy, S., Rastogi, R., and Shim, K. (2000). Efficient algorithms for mining outliers from large data sets. *SIGMOD Rec.*, 29(2):427–438. – Cité pages 39 et 122.

Raparthi, e. a. M. (2023). Predictive Maintenance in IoT Devices using Time Series Analysis and Deep Learning. *Dandao Xuebao / Journal of Ballistics*, 35(3):01–10. – Cité page 52.

Reddy, D. K. K., Behera, H. S., Pratyusha, G. M. S., and Karri, R. (2021). Ensemble Bagging Approach for IoT Sensor Based Anomaly Detection. In Sekhar, G. C., Behera, H. S., Nayak, J., Naik, B., and Pelusi, D., editors, *Intelligent Computing in Control and Communication*, pages 647–665, Singapore. Springer. – Cité page 45.

Refonaa, J., Lakshmi, M., and Vivek, V. (2015). Analysis and prediction of natural disaster using spatial data mining technique. In *2015 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2015]*, pages 1–6. – Cité page 53.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics. – Cité pages 83 et 85.

Riahi, F. and Schulte, O. (2018). Model-based Exception Mining for Object-Relational Data. – Cité page 21.

Ripan, R. C., Sarker, I. H., Hossain, S. M. M., Anwar, M. M., Nowrozy, R., Hoque, M. M., and Furhad, M. H. (2021). A data-driven heart disease prediction model through K-means clustering-based anomaly detection. *SN Computer Science*, 2(2):112. – Cité pages 35 et 40.

Rousseau, P., Camara, D., and Kotzinos, D. (2021). Weak signal detection and identification in large data sets: A review of methods and applications. – Cité page 40.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65. – Cité page 47.

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*, CIKM '17, pages 797–806, New York, NY, USA. Association for Computing Machinery. – Cité page 60.

Ruff, L., Kauffmann, J. R., Vandermeulen, R. A., Montavon, G., Samek, W., Kloft, M., Dietterich, T. G., and Müller, K.-R. (2021). A Unifying Review of Deep and Shallow Anomaly Detection. *Proceedings of the IEEE*, 109(5):756–795. – Cité pages 27, 34, 36, 37 et 41.

Ruff, L., Vandermeulen, R. A., Görnitz, N., Binder, A., Müller, E., Müller, K.-R., and Kloft, M. (2020). Deep Semi-Supervised Anomaly Detection. In *Eighth International Conference on Learning Representations*. – Cité pages 29, 34 et 41.

Ruff, L., Vandermeulen, R. A., Görnitz, N., Deecke, L., Siddiqui, S. A., Binder, A., Müller, E., and Kloft, M. (2018). Deep One-Class Classification. In *Proceedings of Machine Learning Research*, pages 4393–4402. PMLR. – Cité pages 31, 34, 41 et 122.

Ruff, L., Zemlyanskiy, Y., Vandermeulen, R., Schnake, T., and Kloft, M. (2019). Self-Attentive, Multi-Context One-Class Classification for Unsupervised Anomaly Detection on Text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4061–4071, Florence, Italy. Association for Computational Linguistics. – Cité pages 42, 66, 69 et 106.

Ruiz, M. D., Sánchez, D., Delgado, M., and Martin-Bautista, M. J. (2015). Discovering fuzzy exception and anomalous rules. *IEEE Transactions on Fuzzy Systems*, 24(4):930–944. – Cité page 21.

Sabokrou, M., Khalooei, M., Fathy, M., and Adeli, E. (2018). Adversarially Learned One-Class Classifier for Novelty Detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3379–3388, Salt Lake City, UT. IEEE. – Cité page 46.

Safdari, H. and De Bacco, C. (2022). Anomaly detection and community detection in networks. *Journal of Big Data*, 9(1):122. – Cité page 54.

Sakurada, M. and Yairi, T. (2014). Anomaly Detection Using Autoencoders with Nonlinear Dimensionality Reduction. In *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis*, MLSDA'14, pages 4–11, New York, NY, USA. Association for Computing Machinery. – Cité page 42.

Salehi, M., Zhang, X., Bezdek, J. C., and Leckie, C. (2016). Smart Sampling: A Novel Unsupervised Boosting Approach for Outlier Detection. In Kang, B. H. and Bai, Q., editors, *AI 2016: Advances in Artificial Intelligence*, pages 469–481, Cham. Springer International Publishing. – Cité page 45.

Salem, O., Alsubhi, K., Mehaoua, A., and Boutaba, R. (2021). Markov Models for Anomaly Detection in Wireless Body Area Networks for Secure Health Monitoring. *IEEE Journal on Selected Areas in Communications*, 39(2):526–540. – Cité page 52.

Salem, O., Guerassimov, A., Mehaoua, A., Marcus, A., and Furht, B. (2013). Sensor fault and patient anomaly detection and classification in medical wireless sensor networks. In *2013 IEEE International Conference on Communications (ICC)*, pages 4373–4378. – Cité page 20.

Salmon, M., Schumacher, D., and Höhle, M. (2016). Monitoring count time series in R: Aberration detection in public health surveillance. *Journal of Statistical Software*, 70(10):1–35. – Cité pages 20 et 21.

Samariya, D. and Thakkar, A. (2023). A Comprehensive Survey of Anomaly Detection Algorithms. *Annals of Data Science*, 10(3):829–850. – Cité pages 21, 37 et 38.

Saranya, S., Rajeshkumar, R., and Shanthi, S. (2014). A survey on anomaly detection for discovering emerging topics. *International Journal of Computer Science and Mobile Computing*, 310(10):895–902. – Cité page 20.

Sarmadi, H. and Karamodin, A. (2020). A novel anomaly detection method based on adaptive Mahalanobis-squared distance and one-class kNN rule for structural health monitoring under environmental effects. *Mechanical Systems and Signal Processing*, 140:106495. – Cité page 39.

Sarvari, H., Domeniconi, C., Prenkaj, B., and Stilo, G. (2021). Unsupervised Boosting-Based Autoencoder Ensembles for Outlier Detection. In Karlapalem, K., Cheng, H., Ramakrishnan, N., Agrawal, R. K., Reddy, P. K., Srivastava, J., and Chakraborty, T., editors, *Advances in Knowledge Discovery and Data Mining*, pages 91–103, Cham. Springer International Publishing. – Cité pages 36 et 42.

Schlegl, T., Seeböck, P., Waldstein, S. M., Langs, G., and Schmidt-Erfurth, U. (2019). F-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. *Medical Image Analysis*, 54:30–44. – Cité pages 32 et 43.

Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U., and Langs, G. (2017). Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery. In Niethammer, M., Styner, M., Aylward, S., Zhu, H., Oguz, I., Yap, P.-T., and Shen, D., editors, *Information Processing in Medical Imaging*, Lecture Notes in Computer Science, pages 146–157, Cham. Springer International Publishing. – Cité pages 32 et 43.

Schmidl, S., Wenig, P., and Papenbrock, T. (2022). Anomaly detection in time series: A comprehensive evaluation. *Proc. VLDB Endow.*, 15(9):1779–1797. – Cité page 52.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., and Williamson, R. C. (2001). Estimating the Support of a High-Dimensional Distribution. *Neural Computation*, 13(7):1443–1471. – Cité pages 41 et 122.

Schölkopf, B., Williamson, R. C., Smola, A., Shawe-Taylor, J., and Platt, J. (1999). Support vector method for novelty detection. In Solla, S., Leen, T., and Müller, K., editors, *Advances in Neural Information Processing Systems*, volume 12. MIT Press. – Cité pages 21 et 31.

Schubert, E., Zimek, A., and Kriegel, H.-P. (2014). Generalized Outlier Detection with Flexible Kernel Density Estimates. In *Proceedings of the 2014 SIAM International Conference on Data Mining (SDM)*, Proceedings, pages 542–550. Society for Industrial and Applied Mathematics. – Cité page 38.

Schulze, J.-P., Sperl, P., and Böttinger, K. (2022). Anomaly Detection by Recombining Gated Unsupervised Experts. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. – Cité page 45.

Seliya, N., Abdollah Zadeh, A., and Khoshgoftaar, T. M. (2021). A literature review on one-class classification and its potential applications in big data. *Journal of Big Data*, 8(1):122. – Cité page 31.

Seo, S., Seo, D., Jang, M., Jeong, J., and Kang, P. (2020). Unusual customer response identification and visualization based on text mining and anomaly detection. *Expert Systems with Applications*, 144:113111. – Cité page 68.

Shahzad, F., Mannan, A., Javed, A. R., Almadhor, A. S., Baker, T., and Al-Jumeily OBE, D. (2022). Cloud-based multiclass anomaly detection and categorization using ensemble learning. *Journal of Cloud Computing*, 11(1):74. – Cité page 45.

Sharma, S. and Dhama, V. (2020). Abnormal Human Behavior Detection in Video Using Suspicious Object Detection. In Kumar, A., Paprzycki, M., and Gunjan, V. K., editors, *ICDSMLA 2019*, pages 379–388, Singapore. Springer. – Cité page 54.

Shaukat, K., Alam, T. M., Luo, S., Shabbir, S., Hameed, I. A., Li, J., Abbas, S. K., and Javed, U. (2021). A Review of Time-Series Anomaly Detection Techniques: A Step to Future Perspectives. In Arai, K., editor, *Advances in Information and Communication*, pages 865–877, Cham. Springer International Publishing. – Cité page 52.

Shekhar, S., Lu, C.-T., and Zhang, P. (2003). A Unified Approach to Detecting Spatial Outliers. *GeoInformatica*, 7(2):139–166. – Cité page 53.

Sheridan, K., Puranik, T. G., Mangortey, E., Pinon-Fischer, O. J., Kirby, M., and Mavris, D. N. (2020). An application of dbscan clustering for flight anomaly detection during the approach phase. In *AIAA Scitech 2020 Forum*, page 1851. – Cité page 40.

Shewhart, W. A. (1930). Economic Quality Control of Manufactured Product1. *Bell System Technical Journal*, 9(2):364–389. – Cité page 26.

Shi, W., Zhang, L., Li, Y., and Liu, H. (2020). Adversarial semi-supervised learning method for printed circuit board unknown defect detection. *The Journal of Engineering*, 2020(13):505–510. – Cité page 53.

Shin, D.-H., Park, R. C., and Chung, K. (2020). Decision Boundary-Based Anomaly Detection Model Using Improved AnoGAN From ECG Data. *IEEE Access*, 8:108664–108674. – Cité page 43.

Shinnou, H. and Sasaki, M. (2010). Detection of peculiar examples using LOF and one class SVM. In *LREC*. – Cité page 41.

Shu, X., Bao, T., Zhou, Y., Xu, R., Li, Y., and Zhang, K. (2023). Unsupervised dam anomaly detection with spatial–temporal variational autoencoder. *Structural Health Monitoring*, 22(1):39–55. – Cité page 53.

Shylendra, A., Shukla, P., Mukhopadhyay, S., Bhunia, S., and Trivedi, A. R. (2020). Low Power Unsupervised Anomaly Detection by Nonparametric Modeling of Sensor Statistics. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 28(8):1833–1843. – Cité page 38.

Shyu, M.-L., Chen, S.-C., Sarinnapakorn, K., and Chang, L. (2003). A novel anomaly detection scheme based on principal component classifier. Technical report, Miami Univ Coral Gables Fl Dept of Electrical and Computer Engineering. – Cité pages 35, 42 et 122.

Shyu, M.-L., Chen, S.-C., Sarinnapakorn, K., and Chang, L. (2006). Principal Component-based Anomaly Detection Scheme. In Young Lin, T., Ohsuga, S., Liau, C.-J., and Hu, X., editors, *Foundations and Novel Approaches in Data Mining*, pages 311–329. Springer, Berlin, Heidelberg. – Cité page 42.

Simmini, F., Rampazzo, M., Peterle, F., Susto, G. A., and Beghi, A. (2022). A Self-Tuning KPCA-Based Approach to Fault Detection in Chiller Systems. *IEEE Transactions on Control Systems Technology*, 30(4):1359–1374. – Cité page 42.

Singh, A. and Reddy, P. (2024). AnoGAN for Tabular Data: A Novel Approach to Anomaly Detection. – Cité page 43.

Singh, K. V. and Vig, L. (2017). Improved prediction of missing protein interactome links via anomaly detection. *Applied Network Science*, 2(1):2. – Cité page 54.

Singh, M. M. and Kane, N. (2022). Outlier Detection using Ensemble Learning. In *2022 6th International Conference on Information Technology (InCIT)*, pages 234–239. – Cité page 39.

Sivapalan, G., Nundy, K. K., Dev, S., Cardiff, B., and John, D. (2022). ANNet: A Lightweight Neural Network for ECG Anomaly Detection in IoT Edge Sensors. *IEEE Transactions on Biomedical Circuits and Systems*, 16(1):24–35. – Cité page 52.

Siwach, M. and Mann, S. (2022). Anomaly detection for weblog data analysis using weighted PCA technique. *Journal of Information and Optimization Sciences*, 43(1):131–141. – Cité page 35.

Sklar, M. (1959). Fonctions de répartition à n dimensions et leurs marges. In *Annales de l'ISUP*, volume 8, pages 229–231. – Cité page 89.

Smeureanu, S., Ionescu, R. T., Popescu, M., and Alexe, B. (2017). Deep Appearance Features for Abnormal Behavior Detection in Video. In Battiato, S., Gallo, G., Schettini, R., and Stanco, F., editors, *Image Analysis and Processing - ICIAP 2017*, pages 779–789, Cham. Springer International Publishing. – Cité page 54.

Soboroff, I. (2004). Overview of the TREC 2004 Novelty Track. page 16. – Cité page 64.

Soboroff, I. and Harman, D. (2003). Overview of the TREC 2003 Novelty Track. In *TREC*, pages 38–53. Citeseer. – Cité page 64.

Soboroff, I. and Harman, D. (2005). Novelty detection: The TREC experience. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT '05*, pages 105–112, Vancouver, British Columbia, Canada. Association for Computational Linguistics. – Cité page 64.

Solak, M. K. (2009). Detection of multiple outliers in univariate data sets. *Paper SP06-2009, Schering*. – Cité page 37.

Solarz, A., Bilicki, M., Gromadzki, M., Pollo, A., Durkalec, A., and Wypych, M. (2017). Automated novelty detection in the WISE survey with one-class support vector machines. *Astronomy & Astrophysics*, 606:A39. – Cité page 41.

Song, B. and Suh, Y. (2019). Narrative texts-based anomaly detection using accident report documents: The case of chemical process safety. *Journal of Loss Prevention in the Process Industries*, 57:47–54. – Cité pages 40, 60 et 67.

Song, H., Jiang, Z., Men, A., and Yang, B. (2017). A Hybrid Semi-Supervised Anomaly Detection Model for High-Dimensional Data. *Computational Intelligence and Neuroscience*, 2017(1):8501683. – Cité pages 31 et 39.

Srinivasan, R., Wang, L., and Bulleid, J. (2020). Machine learning-based climate time series anomaly detection using convolutional neural networks. *Weather and Climate*, 40(1):16–31. – Cité page 52.

Srivastava, A. and Zane-Ulman, B. (2005). Discovering recurring anomalies in text reports regarding complex space systems. In *2005 IEEE Aerospace Conference*, pages 3853–3862, Big Sky, MT, USA. IEEE. – Cité page 60.

Straub, D., Papaioannou, I., and Betz, W. (2016). Bayesian analysis of rare events. *Journal of Computational Physics*, 314:538–556. – Cité page 21.

Su, C.-R., Hajiyev, J., Fu, C. J., Kao, K.-C., Chang, C.-H., and Chang, C.-T. (2019). A novel framework for a remote patient monitoring (RPM) system with abnormality detection. *Health Policy and Technology*, 8(2):157–170. – Cité page 20.

Su, J., Jiang, C., Jin, X., Qiao, Y., Xiao, T., Ma, H., Wei, R., Jing, Z., Xu, J., and Lin, J. (2024). Large Language Models for Forecasting and Anomaly Detection: A Systematic Literature Review. – Cité pages 21 et 72.

Subudhi, S. and Panigrahi, S. (2015). Quarter-sphere support vector machine for fraud detection in mobile telecommunication networks. *Procedia Computer Science*, 48:353–359. – Cité page 28.

Subudhi, S. and Panigrahi, S. (2022). Application of OPTICS and ensemble learning for database intrusion detection. *Journal of king saud university-computer and information sciences*, 34(3):972–981. – Cité page 40.

Sufi, F. K. (2022). AI-GlobalEvents: A Software for analyzing, identifying and explaining global events with Artificial Intelligence. *Software Impacts*, 11:100218. – Cité pages 20, 60 et 70.

Sundqvist, T., Bhuyan, M. H., Forsman, J., and Elmroth, E. (2020). Boosted Ensemble Learning for Anomaly Detection in 5G RAN. In Maglogiannis, I., Iliadis, L., and Pimenidis, E., editors, *Artificial Intelligence Applications and Innovations*, pages 15–30, Cham. Springer International Publishing. – Cité page 45.

Susto, G. A., Terzi, M., and Beghi, A. (2017). Anomaly Detection Approaches for Semiconductor Manufacturing. *Procedia Manufacturing*, 11:2018–2024. – Cité page 40.

Suzuki, E., Watanabe, T., Yokoi, H., and Takabayashi, K. (2003). Detecting interesting exceptions from medical test data with visual summarization. In *Third IEEE International Conference on Data Mining*, pages 315–322. IEEE. – Cité page 21.

Tack, J., Mo, S., Jeong, J., and Shin, J. (2020). CSI: Novelty detection via contrastive learning on distributionally shifted instances. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 11839–11852. Curran Associates, Inc. – Cité page 21.

Taheri Sarteshnizi, I., Bagloee, S. A., Sarvi, M., and Nassir, N. (2024). Traffic Anomaly Detection: Exploiting Temporal Positioning of Flow-Density Samples. *IEEE Transactions on Intelligent Transportation Systems*, 25(5):4166–4180. – Cité page 40.

Takahashi, T., Tomioka, R., and Yamanishi, K. (2014). Discovering Emerging Topics in Social Streams via Link-Anomaly Detection. *IEEE Transactions on Knowledge and Data Engineering*, 26(1):120–130. – Cité pages 20 et 60.

Takeuchi, J.-i. and Yamanishi, K. (2006). A unifying framework for detecting outliers and change points from time series. *IEEE transactions on Knowledge and Data Engineering*, 18(4):482–492. – Cité page 21.

Tallboys, J., Zhu, Y., and Rajasegarar, S. (2022). Identification of Stock Market Manipulation with Deep Learning. In Li, B., Yue, L., Jiang, J., Chen, W., Li, X., Long, G., Fang, F., and Yu, H., editors, *Advanced Data Mining and Applications*, pages 408–420, Cham. Springer International Publishing. – Cité page 52.

Tang, B. and He, H. (2017). A local density-based approach for outlier detection. *Neurocomputing*, 241:171–180. – Cité pages 38 et 39.

Tang, J., Chen, Z., Fu, A. W.-c., and Cheung, D. W. (2002). Enhancing effectiveness of outlier detections for low density patterns. In Chen, M.-S., Yu, P. S., and Liu, B., editors, *Advances in Knowledge Discovery and Data Mining*, pages 535–548, Berlin, Heidelberg. Springer Berlin Heidelberg. – Cité page 40.

Tang, X., Astle, Y. S., and Freeman, C. (2020). Deep Anomaly Detection with Ensemble-Based Active Learning. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 1663–1670. – Cité page 36.

Tao, X.-M., Chen, W.-H., Du, B.-X., Xu, Y., and Dong, H.-G. (2007). A Novel Model of one-class Bearing Fault Detection using SVDD and Genetic Algorithm. In *2007 2nd IEEE Conference on Industrial Electronics and Applications*, pages 802–807. – Cité page 41.

Tax, D. (2001). *One-Class Classification. Concept Learning in the Absence of Counter Examples Ph. D.* PhD thesis, Thesis (Delft University of Technology, Delft, Netherlands, 2001). – Cité page 31.

Tax, D. M. and Duin, R. P. (2004). Support Vector Data Description. *Machine Learning*, 54(1):45–66. – Cité page 41.

Tayeh, T., Aburakhia, S., Myers, R., and Shami, A. (2022). An Attention-Based ConvLSTM Autoencoder with Dynamic Thresholding for Unsupervised Anomaly Detection in Multivariate Time Series. *Machine Learning and Knowledge Extraction*, 4(2):350–370. – Cité page 26.

Team, G., Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., Silver, D., Johnson, M., Antonoglou, I., Schrittwieser, J., Glaese, A., Chen, J., Pitler, E., Lillicrap, T., Lazaridou, A., Firat, O., Molloy, J., Isard, M., Barham, P. R., Hennigan, T., Lee, B., Viola, F., Reynolds, M., Xu, Y., Doherty, R., Collins, E., Meyer, C., Rutherford, E., Moreira, E., Ayoub, K., Goel, M., Krawczyk, J., Du, C., Chi, E., Cheng, H.-T., Ni, E., Shah, P., Kane, P., Chan, B., Faruqui, M., Severyn, A., Lin, H., Li, Y., Cheng, Y., Ittycheriah, A., Mahdieh, M., Chen, M., Sun, P., Tran, D., Bagri, S., Lakshminarayanan, B., Liu, J., Orban, A., Güra, F., Zhou, H., Song, X., Boffy, A., Ganapathy, H., Zheng, S., Choe, H., Weisz, Á., Zhu, T., Lu, Y., Gopal, S., Kahn, J., Kula, M., Pitman, J., Shah, R., Taropa, E., Merey, M. A., Baeuml, M., Chen, Z., Shafey, L. E., Zhang, Y., Sercinoglu, O., Tucker, G., Piqueras, E., Krikun, M., Barr, I., Savinov, N., Danihelka, I., Roelofs, B., White, A., Andreassen, A., von Glehn, T., Yagati, L., Kazemi, M., Gonzalez, L., Khalman, M., Sygnowski, J., Frechette, A., Smith, C., Culp, L., Proleev, L., Luan, Y., Chen, X., Lottes, J., Schucher, N., Lebron, F., Rrustemi, A., Clay, N., Crone, P., Kocisky, T., Zhao, J., Perz, B., Yu, D., Howard, H., Bloniarz, A., Rae, J. W., Lu, H., Sifre, L., Maggioni, M., Alcober, F., Garrette, D., Barnes, M., Thakoor, S., Austin, J., Barth-Maron, G., Wong, W., Joshi, R., Chaabouni, R., Fatiha, D., Ahuja, A., Tomar, G. S., Senter, E., Chadwick, M., Kornakov, I., Attaluri, N., Iturrate, I., Liu, R., Li, Y., Cogan, S., Chen, J., Jia, C., Gu, C., Zhang, Q., Grimstad, J., Hartman, A. J., Garcia, X., Pillai, T. S., Devlin, J., Laskin, M., Casas, D. d. L., Valter, D., Tao, C., Blanco, L., Badia, A. P., Reitter, D., Chen, M., Brennan, J., Rivera, C., Brin, S., Iqbal, S., Surita, G., Labanowski, J., Rao, A., Winkler, S., Parisotto, E., Gu, Y., Olszewska, K., Addanki, R., Miech, A., Louis, A., Teplyashin, D., Brown, G., Catt, E., Balaguer, J., Xiang, J., Wang, P., Ashwood, Z., Briukhov, A., Webson, A., Ganapathy, S., Sanghavi, S., Kannan, A., Chang, M.-W., Stjerngren, A., Djolonga, J., Sun, Y., Bapna, A., Aitchison, M., Pejman, P., Michalewski, H., Yu, T., Wang, C., Love, J., Ahn, J., Bloxwich, D., Han, K., Humphreys, P., Sellam, T., Bradbury, J., Godbole, V., Samangooei, S., Damoc, B., Kaskasoli, A., Arnold, S. M. R., Vasudevan, V., Agrawal, S., Riesa, J., Lepikhin, D., Tanburn, R., Srinivasan, S., Lim, H., Hodkinson, S.,

Shyam, P., Ferret, J., Hand, S., Garg, A., Paine, T. L., Li, J., Li, Y., Giang, M., Neitz, A., Abbas, Z., York, S., Reid, M., Cole, E., Chowdhery, A., Das, D., Rogozińska, D., Nikolaev, V., Sprechmann, P., Nado, Z., Zilka, L., Prost, F., He, L., Monteiro, M., Mishra, G., Welty, C., Newlan, J., Jia, D., Allamanis, M., Hu, C. H., de Liedekerke, R., Gilmer, J., Saroufim, C., Rijhwani, S., Hou, S., Shrivastava, D., Baddepudi, A., Goldin, A., Ozturel, A., Cassirer, A., Xu, Y., Sohn, D., Sachan, D., Amplayo, R. K., Swanson, C., Petrova, D., Narayan, S., Guez, A., Brahma, S., Landon, J., Patel, M., Zhao, R., Villela, K., Wang, L., Jia, W., Rahtz, M., Giménez, M., Yeung, L., Keeling, J., Georgiev, P., Mincu, D., Wu, B., Haykal, S., Saputro, R., Vodrahalli, K., Qin, J., Cankara, Z., Sharma, A., Fernando, N., Hawkins, W., Neyshabur, B., Kim, S., Hutter, A., Agrawal, P., Castro-Ros, A., van den Driessche, G., Wang, T., Yang, F., Chang, S.-y., Komarek, P., McIlroy, R., Lučić, M., Zhang, G., Farhan, W., Sharman, M., Natsev, P., Michel, P., Bansal, Y., Qiao, S., Cao, K., Shakeri, S., Butterfield, C., Chung, J., Rubenstein, P. K., Agrawal, S., Mensch, A., Soparkar, K., Lenc, K., Chung, T., Pope, A., Maggiore, L., Kay, J., Jhakra, P., Wang, S., Maynez, J., Phuong, M., Tobin, T., Tacchetti, A., Trebacz, M., Robinson, K., Katariya, Y., Riedel, S., Bailey, P., Xiao, K., Ghelani, N., Aroyo, L., Slone, A., Houlsby, N., Xiong, X., Yang, Z., Gribovskaya, E., Adler, J., Wirth, M., Lee, L., Li, M., Kagohara, T., Pavagadhi, J., Bridgers, S., Bortsova, A., Ghemawat, S., Ahmed, Z., Liu, T., Powell, R., Bolina, V., Iinuma, M., Zablotskaia, P., Besley, J., Chung, D.-W., Dozat, T., Comanescu, R., Si, X., Greer, J., Su, G., Polacek, M., Kaufman, R. L., Tokumine, S., Hu, H., Buchatskaya, E., Miao, Y., Elhawaty, M., Siddhant, A., Tomasev, N., Xing, J., Greer, C., Miller, H., Ashraf, S., Roy, A., Zhang, Z., Ma, A., Filos, A., Besta, M., Blevins, R., Klimenko, T., Yeh, C.-K., Changpinyo, S., Mu, J., Chang, O., Pajarskas, M., Muir, C., Cohen, V., Lan, C. L., Haridasan, K., Marathe, A., Hansen, S., Douglas, S., Samuel, R., Wang, M., Austin, S., Lan, C., Jiang, J., Chiu, J., Lorenzo, J. A., Sjösund, L. L., Cevey, S., Gleicher, Z., Avrahami, T., Boral, A., Srinivasan, H., Selo, V., May, R., Aisopos, K., Hussenot, L., Soares, L. B., Baumli, K., Chang, M. B., Recasens, A., Caine, B., Pritzel, A., Pavetic, F., Pardo, F., Gergely, A., Frye, J., Ramasesh, V., Horgan, D., Badola, K., Kassner, N., Roy, S., Dyer, E., Campos, V. C., Tomala, A., Tang, Y., Badawy, D. E., White, E., Mustafa, B., Lang, O., Jindal, A., Vikram, S., Gong, Z., Caelles, S., Hemsley, R., Thornton, G., Feng, F., Stokowiec, W., Zheng, C., Thacker, P., Ünlü, Ç., Zhang, Z., Saleh, M., Svensson, J., Bileschi, M., Patil, P., Anand, A., Ring, R., Tsihlas, K., Vezer, A., Selvi, M., Shevlane, T., Rodriguez, M., Kwiatkowski, T., Daruki, S., Rong, K., Dafoe, A., FitzGerald, N., Gu-Lemberg, K., Khan, M., Hendricks, L. A., Pellat, M., Feinberg, V., Cobon-Kerr, J., Sainath, T., Rauh, M., Hashemi, S. H., Ives, R., Hasson, Y., Noland, E., Cao, Y., Byrd, N., Hou, L., Wang, Q., Sottiaux, T., Paganini, M., Lespiau, J.-B., Moufarek, A., Hassan, S., Shivakumar, K., van Amersfoort, J., Mandhane, A., Joshi, P., Goyal, A., Tung, M., Brock, A., Sheahan, H., Misra, V., Li, C., Rakićević, N., Dehghani, M., Liu, F., Mittal, S., Oh, J., Noury, S., Sezener, E., Huot, F., Lamm, M., De Cao, N., Chen, C., Mudgal, S., Stella, R., Brooks, K., Vasudevan, G., Liu, C., Chain, M., Melinkeri, N., Cohen, A., Wang, V., Seymore, K., Zubkov, S., Goel, R., Yue, S., Krishnakumaran, S., Albert, B., Hurley, N., Sano, M., Mohananey, A., Joughin, J., Filonov, E., Kępa, T., Eldawy, Y., Lim, J., Rishi, R., Badiezadegan, S., Bos, T., Chang, J., Jain, S., Padmanabhan, S. G. S., Puttagunta, S., Krishna, K., Baker, L., Kalb, N., Bedapudi, V., Kurzrok, A., Lei, S., Yu, A., Litvin, O., Zhou, X., Wu, Z., Sobell, S., Siciliano, A., Papir, A., Neale, R., Bragagnolo, J., Toor, T., Chen, T., Anklin, V., Wang, F., Feng, R., Gholami, M., Ling, K., Liu, L., Walter, J., Moghaddam, H.,

Kishore, A., Adamek, J., Mercado, T., Mallinson, J., Wandekar, S., Cagle, S., Ofek, E., Garrido, G., Lombriser, C., Mukha, M., Sun, B., Mohammad, H. R., Matak, J., Qian, Y., Peswani, V., Janus, P., Yuan, Q., Schelin, L., David, O., Garg, A., He, Y., Duzhyi, O., Älgmyr, A., Lottaz, T., Li, Q., Yadav, V., Xu, L., Chinien, A., Shivanna, R., Chuklin, A., Li, J., Spadine, C., Wolfe, T., Mohamed, K., Das, S., Dai, Z., He, K., von Dincklage, D., Upadhyay, S., Maurya, A., Chi, L., Krause, S., Salama, K., Rabinovitch, P. G., M, P. K. R., Selvan, A., Dektiarev, M., Ghiasi, G., Guven, E., Gupta, H., Liu, B., Sharma, D., Shtacher, I. H., Paul, S., Akerlund, O., Aubet, F.-X., Huang, T., Zhu, C., Zhu, E., Teixeira, E., Fritze, M., Bertolini, F., Marinescu, L.-E., Bölle, M., Paulus, D., Gupta, K., Latkar, T., Chang, M., Sanders, J., Wilson, R., Wu, X., Tan, Y.-X., Thiet, L. N., Doshi, T., Lall, S., Mishra, S., Chen, W., Luong, T., Benjamin, S., Lee, J., Andrejczuk, E., Rabiej, D., Ranjan, V., Styrc, K., Yin, P., Simon, J., Harriott, M. R., Bansal, M., Robsky, A., Bacon, G., Greene, D., Mirylenka, D., Zhou, C., Sarvana, O., Goyal, A., Andermatt, S., Siegler, P., Horn, B., Israel, A., Pongetti, F., Chen, C.-W. L., Selvatici, M., Silva, P., Wang, K., Tolins, J., Guu, K., Yogev, R., Cai, X., Agostini, A., Shah, M., Nguyen, H., Donnaile, N. Ó., Pereira, S., Friso, L., Stambler, A., Kurzrok, A., Kuang, C., Romanikhin, Y., Geller, M., Yan, Z. J., Jang, K., Lee, C.-C., Fica, W., Malmi, E., Tan, Q., Banica, D., Balle, D., Pham, R., Huang, Y., Avram, D., Shi, H., Singh, J., Hidey, C., Ahuja, N., Saxena, P., Dooley, D., Potharaju, S. P., O'Neill, E., Gokulchandran, A., Foley, R., Zhao, K., Dusenberry, M., Liu, Y., Mehta, P., Kotikalapudi, R., Safranek-Shrader, C., Goodman, A., Kessinger, J., Globen, E., Kolhar, P., Gorgolewski, C., Ibrahim, A., Song, Y., Eichenbaum, A., Brovelli, T., Potluri, S., Lahoti, P., Baetu, C., Ghorbani, A., Chen, C., Crawford, A., Pal, S., Sridhar, M., Gurita, P., Mujika, A., Petrovski, I., Cedoz, P.-L., Li, C., Chen, S., Santo, N. D., Goyal, S., Punjabi, J., Kappaganthu, K., Kwak, C., LV, P., Velury, S., Choudhury, H., Hall, J., Shah, P., Figueira, R., Thomas, M., Lu, M., Zhou, T., Kumar, C., Jurdi, T., Chikkerur, S., Ma, Y., Yu, A., Kwak, S., Ähdel, V., Rajayogam, S., Choma, T., Liu, F., Barua, A., Ji, C., Park, J. H., Hellendoorn, V., Bailey, A., Bilal, T., Zhou, H., Khatir, M., Sutton, C., Rzadkowski, W., Macintosh, F., Shagin, K., Medina, P., Liang, C., Zhou, J., Shah, P., Bi, Y., Dankovics, A., Banga, S., Lehmann, S., Bredesen, M., Lin, Z., Hoffmann, J. E., Lai, J., Chung, R., Yang, K., Balani, N., Bražinskas, A., Sozanschi, A., Hayes, M., Alcalde, H. F., Makarov, P., Chen, W., Stella, A., Snijders, L., Mandl, M., Kärrman, A., Nowak, P., Wu, X., Dyck, A., Vaidyanathan, K., R, R., Mallet, J., Rudominer, M., Johnston, E., Mittal, S., Udathu, A., Christensen, J., Verma, V., Irving, Z., Santucci, A., Elsayed, G., Davoodi, E., Georgiev, M., Tenney, I., Hua, N., Cideron, G., Leurent, E., Alnahlawi, M., Georgescu, I., Wei, N., Zheng, I., Scandinaro, D., Jiang, H., Snoek, J., Sundararajan, M., Wang, X., Ontiveros, Z., Karo, I., Cole, J., Rajashekhar, V., Tumeh, L., Ben-David, E., Jain, R., Uesato, J., Datta, R., Bunyan, O., Wu, S., Zhang, J., Stanczyk, P., Zhang, Y., Steiner, D., Naskar, S., Azzam, M., Johnson, M., Paszke, A., Chiu, C.-C., Elias, J. S., Mohiuddin, A., Muhammad, F., Miao, J., Lee, A., Vieillard, N., Park, J., Zhang, J., Stanway, J., Garmon, D., Karmarkar, A., Dong, Z., Lee, J., Kumar, A., Zhou, L., Evens, J., Isaac, W., Irving, G., Loper, E., Fink, M., Arkatkar, I., Chen, N., Shafran, I., Petrychenko, I., Chen, Z., Jia, J., Levskaya, A., Zhu, Z., Grabowski, P., Mao, Y., Magni, A., Yao, K., Snaider, J., Casagrande, N., Palmer, E., Suganthan, P., Castaño, A., Giannoumis, I., Kim, W., Rybiński, M., Sreevatsa, A., Prendki, J., Soergel, D., Goedeckemeyer, A., Gierke, W., Jafari, M., Gaba, M., Wiesner, J., Wright, D. G., Wei, Y., Vashisht, H., Kulizhskaya, Y., Hoover, J., Le, M., Li, L., Iwuanyanwu, C., Liu, L., Ramirez, K., Khorlin, A., Cui, A., LIN,

T., Wu, M., Aguilar, R., Pallo, K., Chakladar, A., Perng, G., Abellan, E. A., Zhang, M., Dasgupta, I., Kushman, N., Penchev, I., Repina, A., Wu, X., van der Weide, T., Ponnapalli, P., Kaplan, C., Simsa, J., Li, S., Dousse, O., Yang, F., Piper, J., Ie, N., Pasumarthi, R., Lintz, N., Vijayakumar, A., Andor, D., Valenzuela, P., Lui, M., Paduraru, C., Peng, D., Lee, K., Zhang, S., Greene, S., Nguyen, D. D., Kurylowicz, P., Hardin, C., Dixon, L., Janzer, L., Choo, K., Feng, Z., Zhang, B., Singhal, A., Du, D., McKinnon, D., Antropova, N., Bolukbasi, T., Keller, O., Reid, D., Finchelstein, D., Raad, M. A., Crocker, R., Hawkins, P., Dadashi, R., Gaffney, C., Franko, K., Bulanova, A., Leblond, R., Chung, S., Askham, H., Cobo, L. C., Xu, K., Fischer, F., Xu, J., Sorokin, C., Alberti, C., Lin, C.-C., Evans, C., Dimitriev, A., Forbes, H., Banarse, D., Tung, Z., Omernick, M., Bishop, C., Sterneck, R., Jain, R., Xia, J., Amid, E., Piccinno, F., Wang, X., Banzal, P., Mankowitz, D. J., Polozov, A., Krakovna, V., Brown, S., Bateni, M., Duan, D., Firoiu, V., Thotakuri, M., Natan, T., Geist, M., tan Girgin, S., Li, H., Ye, J., Roval, O., Tojo, R., Kwong, M., Lee-Thorp, J., Yew, C., Sinopalnikov, D., Ramos, S., Mellor, J., Sharma, A., Wu, K., Miller, D., Sonnerat, N., Vnukov, D., Greig, R., Beattie, J., Caveness, E., Bai, L., Eisenschlos, J., Korchemniy, A., Tsai, T., Jasarevic, M., Kong, W., Dao, P., Zheng, Z., Liu, F., Yang, F., Zhu, R., Teh, T. H., Sanmiya, J., Gladchenko, E., Trdin, N., Toyama, D., Rosen, E., Tavakkol, S., Xue, L., Elkind, C., Woodman, O., Carpenter, J., Papamakarios, G., Kemp, R., Kafle, S., Grunina, T., Sinha, R., Talbert, A., Wu, D., Owusu-Afriyie, D., Du, C., Thornton, C., Pont-Tuset, J., Narayana, P., Li, J., Fatehi, S., Wieting, J., Ajmeri, O., Uria, B., Ko, Y., Knight, L., Héliou, A., Niu, N., Gu, S., Pang, C., Li, Y., Levine, N., Stolovich, A., Santamaria-Fernandez, R., Goenka, S., Yustalim, W., Strudel, R., Elqursh, A., Deck, C., Lee, H., Li, Z., Levin, K., Hoffmann, R., Holtmann-Rice, D., Bachem, O., Arora, S., Koh, C., Yeganeh, S. H., Põder, S., Tariq, M., Sun, Y., Ionita, L., Seyedhosseini, M., Tafti, P., Liu, Z., Gulati, A., Liu, J., Ye, X., Chrzaszcz, B., Wang, L., Sethi, N., Li, T., Brown, B., Singh, S., Fan, W., Parisi, A., Stanton, J., Koverkathu, V., Choquette-Choo, C. A., Li, Y., Lu, T. J., Ittycheriah, A., Shroff, P., Varadarajan, M., Bahargam, S., Willoughby, R., Gaddy, D., Desjardins, G., Cornero, M., Robenek, B., Mittal, B., Albrecht, B., Shenoy, A., Moiseev, F., Jacobsson, H., Ghaffarkhah, A., Rivière, M., Walton, A., Crepy, C., Parrish, A., Zhou, Z., Farabet, C., Radebaugh, C., Srinivasan, P., van der Salm, C., Fidjeland, A., Scellato, S., Latorre-Chimoto, E., Klimczak-Plucińska, H., Bridson, D., de Cesare, D., Hudson, T., Mendolicchio, P., Walker, L., Morris, A., Mauger, M., Guseynov, A., Reid, A., Odoom, S., Loher, L., Cotruta, V., Yenugula, M., Grewe, D., Petrushkina, A., Duerig, T., Sanchez, A., Yadlowsky, S., Shen, A., Globerson, A., Webb, L., Dua, S., Li, D., Bhupatiraju, S., Hurt, D., Qureshi, H., Agarwal, A., Shani, T., Eyal, M., Khare, A., Belle, S. R., Wang, L., Tekur, C., Kale, M. S., Wei, J., Sang, R., Saeta, B., Liechty, T., Sun, Y., Zhao, Y., Lee, S., Nayak, P., Fritz, D., Vuyyuru, M. R., Aslanides, J., Vyas, N., Wicke, M., Ma, X., Eltyshev, E., Martin, N., Cate, H., Manyika, J., Amiri, K., Kim, Y., Xiong, X., Kang, K., Luisier, F., Tripuraneni, N., Madras, D., Guo, M., Waters, A., Wang, O., Ainslie, J., Baldridge, J., Zhang, H., Pruthi, G., Bauer, J., Yang, F., Mansour, R., Gelman, J., Xu, Y., Polovets, G., Liu, J., Cai, H., Chen, W., Sheng, X., Xue, E., Ozair, S., Angermueller, C., Li, X., Sinha, A., Wang, W., Wiesinger, J., Koukoumidis, E., Tian, Y., Iyer, A., Gurumurthy, M., Goldenson, M., Shah, P., Blake, M. K., Yu, H., Urbanowicz, A., Palomaki, J., Fernando, C., Durden, K., Mehta, H., Momchev, N., Rahimtoroghi, E., Georgaki, M., Raul, A., Ruder, S., Redshaw, M., Lee, J., Zhou, D., Jalan, K., Li, D., Hechtman, B., Schuh, P., Nasr, M., Milan, K., Mikulik, V., Franco, J., Green, T., Nguyen, N., Kelley, J., Mahendru, A., Hu, A., Howland, J., Vargas, B.,

Hui, J., Bansal, K., Rao, V., Ghiya, R., Wang, E., Ye, K., Sarr, J. M., Preston, M. M., Elish, M., Li, S., Kaku, A., Gupta, J., Pasupat, I., Juan, D.-C., Someswar, M., M., T., Chen, X., Amini, A., Fabrikant, A., Chu, E., Dong, X., Muthal, A., Buthpitiya, S., Jauhari, S., Hua, N., Khandelwal, U., Hitron, A., Ren, J., Rinaldi, L., Drath, S., Dabush, A., Jiang, N.-J., Godhia, H., Sachs, U., Chen, A., Fan, Y., Taitelbaum, H., Noga, H., Dai, Z., Wang, J., Liang, C., Hamer, J., Ferng, C.-S., Elkind, C., Atias, A., Lee, P., Listík, V., Carlen, M., van de Kerkhof, J., Pikus, M., Zaher, K., Müller, P., Zykova, S., Stefanec, R., Gatsko, V., Hirnschall, C., Sethi, A., Xu, X. F., Ahuja, C., Tsai, B., Stefanoiu, A., Feng, B., Dhandhania, K., Katyal, M., Gupta, A., Parulekar, A., Pitta, D., Zhao, J., Bhatia, V., Bhavnani, Y., Alhadlaq, O., Li, X., Danenberg, P., Tu, D., Pine, A., Filippova, V., Ghosh, A., Limonchik, B., Urala, B., Lanka, C. K., Clive, D., Sun, Y., Li, E., Wu, H., Hongtongsak, K., Li, I., Thakkar, K., Omarov, K., Majmundar, K., Alverson, M., Kucharski, M., Patel, M., Jain, M., Zabelin, M., Pelagatti, P., Kohli, R., Kumar, S., Kim, J., Sankar, S., Shah, V., Ramachandruni, L., Zeng, X., Bariach, B., Weidinger, L., Vu, T., Subramanya, A., Hsiao, S., Hassabis, D., Kavukcuoglu, K., Sadovsky, A., Le, Q., Strohman, T., Wu, Y., Petrov, S., Dean, J., and Vinyals, O. (2024). Gemini: A Family of Highly Capable Multimodal Models. – Cité page 171.

Ten, C.-W., Hong, J., and Liu, C.-C. (2011). Anomaly Detection for Cybersecurity of the Substations. *IEEE Transactions on Smart Grid*, 2(4):865–873. – Cité page 20.

Thang, T. M. and Kim, J. (2011). The Anomaly Detection by Using DBSCAN Clustering with Multiple Parameters. In *2011 International Conference on Information Science and Applications*, pages 1–5. – Cité page 40.

Theofilatos, A., Yannis, G., Kopelias, P., and Papadimitriou, F. (2019). Impact of real-time traffic characteristics on crash occurrence: Preliminary results of the case of rare events. *Accident Analysis & Prevention*, 130:151–159. – Cité page 21.

Thuc, H. L. U., Tuan, P. V., and Hwang, J.-N. (2017). An effective video-based model for fall monitoring of the elderly. In *2017 International Conference on System Science and Engineering (ICSSE)*, pages 48–52. – Cité page 54.

Thudumu, S., Branch, P., Jin, J., and Singh, J. (2020). A comprehensive survey of anomaly detection techniques for high dimensional big data. *Journal of Big Data*, 7(1):42. – Cité pages 38 et 42.

Todkar, S. S., Baltazart, V., Ihamouten, A., Dérobert, X., and Guilbert, D. (2021). One-class SVM based outlier detection strategy to detect thin interlayer debondings within pavement structures using Ground Penetrating Radar data. *Journal of Applied Geophysics*, 192:104392. – Cité page 41.

Tonnaer, L., Li, J., Osin, V., Holenderski, M., and Menkovski, V. (2019). Anomaly Detection for Visual Quality Control of 3D-Printed Products. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. – Cité page 53.

Torshizi, A. S. and Ghazikhani, A. (2019). Automatic Twitter Rumor Detection Based on LSTM Classifier. In Grandinetti, L., Mirtaheri, S. L., and Shahbazian, R., editors, *High-Performance Computing and Big Data Analysis*, pages 291–300, Cham. Springer International Publishing. – Cité page 60.

Toth, E. and Chawla, S. (2018). Group Deviation Detection Methods: A Survey. *ACM Computing Surveys*, 51(4):1–38. – Cité page 21.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023a). LLaMA: Open and Efficient Foundation Language Models. – Cité pages 169 et 173.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. (2023b). Llama 2: Open Foundation and Fine-Tuned Chat Models. – Cité page 170.

Tran, L., Fan, L., and Shahabi, C. (2016). Distance-based outlier detection in data streams. *Proc. VLDB Endow.*, 9(12):1089–1100. – Cité page 39.

Tran, P. H., Tran, K. P., Huong, T. T., Heuchenne, C., HienTran, P., and Le, T. M. H. (2018). Real Time Data-Driven Approaches for Credit Card Fraud Detection. In *Proceedings of the 2018 International Conference on E-Business and Applications*, ICEBA 2018, pages 6–9, New York, NY, USA. Association for Computing Machinery. – Cité page 52.

Tsai, F. S. (2010). *Techniques for Intelligent Novelty Mining*. PhD thesis, Nanyang Technological University. – Cité page 108.

Tsogbaatar, E., Bhuyan, M. H., Taenaka, Y., Fall, D., Gonchigsumlaa, K., Elmroth, E., and Kadobayashi, Y. (2021). DeL-IoT: A deep ensemble learning approach to uncover anomalies in IoT. *Internet of Things*, 14:100391. – Cité page 36.

Ullah, I. and Mahmoud, Q. H. (2021). Design and Development of a Deep Learning-Based Model for Anomaly Detection in IoT Networks. *IEEE Access*, 9:103906–103926. – Cité page 20.

Urvoy, M. and Autrusseau, F. (2014). Application of Grubbs' test for outliers to the detection of watermarks. In *Proceedings of the 2nd ACM Workshop on Information Hiding and Multimedia Security*, IH&MMSec '14, pages 49–60, New York, NY, USA. Association for Computing Machinery. – Cité page 37.

van Hespen, K. M., Zwanenburg, J. J. M., Dankbaar, J. W., Geerlings, M. I., Hendrikse, J., and Kuijf, H. J. (2021). An anomaly detection approach to identify chronic brain infarcts on MRI. *Scientific Reports*, 11(1):7714. – Cité page 53.

Vartouni, A. M., Kashi, S. S., and Teshnehlab, M. (2018). An anomaly detection method to detect web attacks using Stacked Auto-Encoder. In *2018 6th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS)*, pages 131–134. – Cité page 20.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc. – Cité page 162.

Verma, A. and Maiti, J. (2018). Text-document clustering-based cause and effect analysis methodology for steel plant incident data. *International Journal of Injury Control and Safety Promotion*, 25(4):416–426. – Cité page 60.

Villa-Pérez, M. E., Álvarez-Carmona, M. Á., Loyola-González, O., Medina-Pérez, M. A., Velazco-Rossell, J. C., and Choo, K.-K. R. (2021). Semi-supervised anomaly detection algorithms: A comparative summary and future research directions. *Knowledge-Based Systems*, 218:106878. – Cité pages 29, 31 et 45.

Volkau, I., Mujeeb, A., Wenting, D., Marius, E., and Alexei, S. (2019). Detection Defect in Printed Circuit Boards using Unsupervised Feature Extraction Upon Transfer Learning. In *2019 International Conference on Cyberworlds (CW)*, pages 101–108. – Cité page 53.

Wagner, A. and Plattner, B. (2005). Entropy based worm and anomaly detection in fast IP networks. In *14th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprise (WETICE'05)*, pages 172–177. – Cité page 44.

Wahl, M. (2021). Detecting Hate Speech in Norwegian Texts Using BERT Semi-Supervised Anomaly Detection. Master's thesis, Norwegian University of Science and Technology. – Cité page 111.

Wahyono, T., Heryadi, Y., Soeparno, H., and Abbas, B. S. (2020). Anomaly Detection in Climate Data Using Stacked and Densely Connected Long Short-Term Memory Model. 31(4). – Cité page 52.

Walambe, R., Marathe, A., Kotecha, K., and Ghinea, G. (2021). Lightweight Object Detection Ensemble Framework for Autonomous Vehicles in Challenging Weather Conditions. *Computational Intelligence and Neuroscience*, 2021(1):5278820. – Cité page 54.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2019a). Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32. – Cité page 108.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018a). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355. – Cité page 108.

Wang, B., Huang, J., Zheng, H., and Wu, H. (2016a). Semi-Supervised Recursive Autoencoders for Social Review Spam Detection. In *2016 12th International Conference on Computational Intelligence and Security (CIS)*, pages 116–119. – Cité page 60.

Wang, G., Yang, J., and Li, R. (2017). Imbalanced SVM-Based Anomaly Detection Algorithm for Imbalanced Training Datasets. *ETRI Journal*, 39(5):621–631. – Cité page 28.

Wang, H., Peng, M.-j., Yu, Y., Saeed, H., Hao, C.-m., and Liu, Y.-k. (2021). Fault identification and diagnosis based on KPCA and similarity clustering for nuclear power plants. *Annals of Nuclear Energy*, 150:107786. – Cité page 42.

Wang, H., Yang, R., and Shi, J. (2023a). Anomaly Detection in Financial Transactions Via Graph-Based Feature Aggregations. In Wrembel, R., Gamper, J., Kotsis, G., Tjoa, A. M., and Khalil, I., editors, *Big Data Analytics and Knowledge Discovery*, pages 64–79, Cham. Springer Nature Switzerland. – Cité page 54.

Wang, J. and Chen, Y.-J. (2019). A novelty detection patent mining approach for analyzing technological opportunities. *Advanced Engineering Informatics*, 42:100941. – Cité page 60.

Wang, J. and Cherian, A. (2019). Gods: Generalized one-class discriminative subspaces for anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8201–8211. – Cité page 31.

Wang, W., Wang, Y., Tan, X., Liu, Y., and Yang, a. S. (2018b). BotCapturer: Detecting Botnets based on Two-Layered Analysis with Graph Anomaly Detection and Network Traffic Clustering. *International Journal of Performability Engineering*, 14(5):1050. – Cité page 54.

Wang, W., Zhang, B., Wang, D., Jiang, Y., Qin, S., and Xue, L. (2016b). Anomaly detection based on probability density function with Kullback–Leibler divergence. *Signal Processing*, 126:12–17. – Cité page 44.

Wang, X.-k., Hou, W.-h., Zhang, H.-y., Wang, J.-q., Goh, M., Tian, Z.-p., and Shen, K.-w. (2022). KDE-OCSVM model using Kullback-Leibler divergence to detect anomalies in medical claims. *Expert Systems with Applications*, 200:117056. – Cité page 39.

Wang, Y., Yu, Z., and Zhu, L. (2023b). Intrusion detection for high-speed railways based on unsupervised anomaly detection models. *Applied Intelligence*, 53(7):8453–8466. – Cité page 54.

Wang, Y., Zhu, S., and Li, C. (2019b). Research on An Ensemble Anomaly Detection Algorithm. *Journal of Physics: Conference Series*, 1314(1):012198. – Cité pages 38 et 39.

Wang, Y. F., Jiong, Y., Su, G. P., and Qian, Y. R. (2019c). A new outlier detection method based on OPTICS. *Sustainable cities and society*, 45:197–212. – Cité page 40.

Wang, Z., Zhou, Y., and Li, G. (2020). Anomaly detection by using streaming K-means and batch K-means. In *2020 5th IEEE International Conference on Big Data Analytics (ICBDA)*, pages 11–17. IEEE. – Cité page 35.

Wayne, C. L. (1997). Topic Detection & Tracking (TDT): Overview & Perspective. In *Workshop Held at the University of Maryland On*, volume 27, page 28. Citeseer. – Cité page 64.

Wazid, M. and Das, A. K. (2016). An efficient hybrid anomaly detection scheme using K-means clustering for wireless sensor networks. *Wireless Personal Communications*, 90:1971–2000. – Cité pages 35 et 40.

Weber, M., Domeniconi, G., Chen, J., Weidele, D. K. I., Bellei, C., Robinson, T., and Leiserson, C. E. (2019). Anti-Money Laundering in Bitcoin: Experimenting with Graph Convolutional Networks for Financial Forensics. – Cité page 20.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. (2022a). Finetuned Language Models Are Zero-Shot Learners. – Cité pages 165 et 172.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., and Fedus, W. (2022b). Emergent Abilities of Large Language Models. – Cité page 164.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. (2023). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. – Cité page 165.

Wittkopp, T., Scheinert, D., Wiesner, P., Acker, A., and Kao, O. (2023). PULL: Reactive Log Anomaly Detection Based On Iterative PU Learning. – Cité page 32.

Wurzer, D., Lavrenko, V., and Osborne, M. (2015). Twitter-scale New Event Detection via K-term Hashing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2584–2589, Lisbon, Portugal. Association for Computational Linguistics. – Cité page 60.

Xia, X., Pan, X., Li, N., He, X., Ma, L., Zhang, X., and Ding, N. (2022). GAN-based anomaly detection: A review. *Neurocomputing*, 493(C):497–535. – Cité page 43.

Xiao, Y., Wang, H., and Xu, W. (2014). Model selection of Gaussian kernel PCA for novelty detection. *Chemometrics and Intelligent Laboratory Systems*, 136:164–172. – Cité page 42.

Xie, M., Hu, J., Guo, S., and Zomaya, A. Y. (2017). Distributed Segment-Based Anomaly Detection With Kullback–Leibler Divergence in Wireless Sensor Networks. *IEEE Transactions on Information Forensics and Security*, 12(1):101–110. – Cité page 44.

Xie, M., Hu, J., Han, S., and Chen, H.-H. (2013). Scalable Hypergrid k-NN-Based Online Anomaly Detection in Wireless Sensor Networks. *IEEE Transactions on Parallel and Distributed Systems*, 24(8):1661–1670. – Cité page 39.

Xie, M., Hu, J., and Tian, B. (2012). Histogram-Based Online Anomaly Detection in Hierarchical Wireless Sensor Networks. In *2012 IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications*, pages 751–759. – Cité page 38.

Xing, H.-J. and Liu, W.-T. (2020). Robust AdaBoost based ensemble of one-class support vector machines. *Information Fusion*, 55:45–58. – Cité page 45.

Xu, A., Ren, X., and Jia, R. (2023a). Contrastive Novelty-Augmented Learning: Anticipating Outliers with Large Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11778–11801, Toronto, Canada. Association for Computational Linguistics. – Cité page 21.

Xu, D., Wang, Y., Meng, Y., and Zhang, Z. (2017). An Improved Data Anomaly Detection Method Based on Isolation Forest. In *2017 10th International Symposium on Computational Intelligence and Design (ISCID)*, pages 287–291, Hangzhou. IEEE. – Cité page 45.

Xu, H., Pang, G., Wang, Y., and Wang, Y. (2023b). Deep Isolation Forest for Anomaly Detection. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12591–12604. – Cité pages 21, 45, 113 et 122.

Xu, H., Zhang, L., Li, P., and Zhu, F. (2022). Outlier detection algorithm based on k-nearest neighbors-local outlier factor. *Journal of Algorithms & Computational Technology*, 16:17483026221078111. – Cité page 40.

Xu, L., Hu, H., Zhang, X., Li, L., Cao, C., Li, Y., Xu, Y., Sun, K., Yu, D., Yu, C., Tian, Y., Dong, Q., Liu, W., Shi, B., Cui, Y., Li, J., Zeng, J., Wang, R., Xie, W., Li, Y., Patterson, Y., Tian, Z., Zhang, Y., Zhou, H., Liu, S., Zhao, Z., Zhao, Q., Yue, C., Zhang, X., Yang, Z., Richardson, K., and Lan, Z. (2020). CLUE: A Chinese Language Understanding Evaluation Benchmark. In Scott, D., Bel, N., and Zong, C., editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772, Barcelona, Spain (Online). International Committee on Computational Linguistics. – Cité page 108.

Xu, S., Wu, H., and Bie, R. (2019a). CXNet-m1: Anomaly Detection on Chest X-Rays With Image-Based Deep Learning. *IEEE Access*, 7:4466–4477. – Cité page 53.

Xu, W., Huang, L., Fox, A., Patterson, D., and Jordan, M. I. (2009). Detecting large-scale system problems by mining console logs. In *Proceedings of the ACM SIGOPS 22nd Symposium on Operating Systems Principles*, pages 117–132, Big Sky Montana USA. ACM. – Cité page 20.

Xu, X., Liu, H., and Yao, M. (2019b). Recent Progress of Anomaly Detection. *Complexity*, 2019:1–11. – Cité pages 45 et 46.

Yamanaka, Y., Iwata, T., Takahashi, H., Yamada, M., and Kanai, S. (2019). Autoencoding Binary Classifiers for Supervised Anomaly Detection. In Nayak, A. C. and Sharma, A., editors, *PRICAI 2019: Trends in Artificial Intelligence*, Lecture Notes in Computer Science, pages 647–659, Cham. Springer International Publishing. – Cité page 28.

Yang, K., Kpotufe, S., and Feamster, N. (2021). An Efficient One-Class SVM for Anomaly Detection in the Internet of Things. *arXiv:2104.11146 [cs]*. – Cité pages 31 et 41.

Yang, Y., Zhang, J., Carbonell, J., and Jin, C. (2002). Topic-conditioned novelty detection. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 688–693. – Cité page 21.

Yao, Y., Xu, M., Wang, Y., Crandall, D. J., and Atkins, E. M. (2019). Unsupervised traffic accident detection in first-person videos. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 273–280. IEEE. – Cité page 54.

Yap, T. Y. (2020). Text Anomaly Detection with ARAE-AnoGAN. *Honors Projects*, 22. – Cité pages 32 et 66.

Ye, F., Zheng, H., Huang, C., and Zhang, Y. (2021). Deep Unsupervised Image Anomaly Detection: An Information Theoretic Framework. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 1609–1613. – Cité page 44.

Ying, S., Wang, B., Wang, L., Li, Q., Zhao, Y., Shang, J., Huang, H., Cheng, G., Yang, Z., and Geng, J. (2021). An Improved KNN-Based Efficient Log Anomaly Detection Method with Automatically Labeled Samples. *ACM Trans. Knowl. Discov. Data*, 15(3):34:1–34:22. – Cité page 39.

Yong, D., Yuanpeng, Z., Yaqing, X., Yu, P., and Datong, L. (2017). Unmanned aerial vehicle sensor data anomaly detection using kernel principle component analysis. In *2017 13th IEEE International Conference on Electronic Measurement & Instruments (ICEMI)*, pages 241–246. – Cité page 42.

Yong, W., Guiyun, M., Xu, C., and Zhengying, W. (2022). Anomaly Detection of Semiconductor Processing Data Based on DTW-LOF Algorithm. In *2022 China Semiconductor Technology International Conference (CSTIC)*, pages 1–3. – Cité pages 35 et 40.

Yoon, J. (2012). Detecting weak signals for long-term business opportunities using text mining of Web news. *Expert Systems with Applications*, 39(16):12543–12550. – Cité page 60.

You, Z., Zhou, Y., Yang, T., and Fan, W. (2021). Anomaly-Injected Deep Support Vector Data Description for Text Outlier Detection. – Cité page 66.

Youssef, A., Delpha, C., and Diallo, D. (2016). An optimal fault detection threshold for early detection using Kullback–Leibler Divergence for unknown distribution data. *Signal Processing*, 120:266–279. – Cité page 44.

Yu, Q., Kavitha, M. S., and Kurita, T. (2021). Mixture of experts with convolutional and variational autoencoders for anomaly detection. *Applied Intelligence*, 51(6):3241–3254. – Cité page 45.

Yu, R., Qiu, H., Wen, Z., Lin, C.-Y., and Liu, Y. (2016). A Survey on Social Media Anomaly Detection. 18(1):14. – Cité page 20.

Yuan, D., Miao, Y., Gong, N. Z., Yang, Z., Li, Q., Song, D., Wang, Q., and Liang, X. (2019a). Detecting Fake Accounts in Online Social Networks at the Time of Registrations. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, CCS '19, pages 1423–1438, New York, NY, USA. Association for Computing Machinery. – Cité page 54.

Yuan, M., Boston-Fisher, N., Luo, Y., Verma, A., and Buckeridge, D. L. (2019b). A systematic review of aberration detection algorithms used in public health surveillance. *Journal of Biomedical Informatics*, 94:103181. – Cité page 21.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019). Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of the 2019 Conference of the North*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics. – Cité page 111.

Zenati, H., Romain, M., Foo, C. S., Lecouat, B., and Chandrasekhar, V. R. (2018). Adversarially Learned Anomaly Detection. – Cité pages 36, 42, 43, 92 et 103.

Zeng, J., Kruger, U., Geluk, J., Wang, X., and Xie, L. (2014). Detecting abnormal situations using the Kullback–Leibler divergence. *Automatica*, 50(11):2777–2786. – Cité page 44.

Zeng, Z., Ni, W., Fang, T., Li, X., Zhao, X., and Song, Y. (2022). Weakly Supervised Text Classification using Supervision Signals from a Language Model. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2295–2305, Seattle, United States. Association for Computational Linguistics. – Cité page 66.

Zeufack, V., Kim, D., Seo, D., and Lee, A. (2021). An unsupervised anomaly detection framework for detecting anomalies in real time through network system's log files analysis. *High-Confidence Computing*, 1(2):100030. – Cité page 40.

Zhang, A., Zhao, X., and Wang, L. (2021a). CNN and LSTM based Encoder-Decoder for Anomaly Detection in Multivariate Time Series. In *2021 IEEE 5th Information Technology,Networking,Electronic and Automation Control Conference (ITNEC)*, volume 5, pages 571–575. – Cité page 52.

Zhang, C. and Zuo, R. (2021). Recognition of multivariate geochemical anomalies associated with mineralization using an improved generative adversarial network. *Ore Geology Reviews*, 136:104264. – Cité page 43.

Zhang, J., Wang, Z., Meng, J., Tan, Y.-P., and Yuan, J. (2019). Boosting Positive and Unlabeled Learning for Anomaly Detection With Multi-Features. *IEEE Transactions on Multimedia*, 21(5):1332–1344. – Cité page 32.

Zhang, J., Wang, Z., Yuan, J., and Tan, Y.-P. (2017). Positive and Unlabeled Learning for Anomaly Detection with Multi-features. In *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17, pages 854–862, New York, NY, USA. Association for Computing Machinery. – Cité page 32.

Zhang, W., Yang, D., Zhang, S., Ablanedo-Rosas, J. H., Wu, X., and Lou, Y. (2021b). A novel multi-stage ensemble model with enhanced outlier adaptation for credit scoring. *Expert Systems with Applications*, 165:113872. – Cité page 45.

Zhang, X., Junbao, Z., and LeCun, Y. (2015). Character-level Convolutional Networks for Text Classification. *Advances in neural information processing systems*, 28(arXiv:1502.01710). – Cité pages 66 et 107.

Zhang, Y., Liang, W., Yuan, X., Zhang, S., Yang, G., and Zeng, Z. (2024). Deep Learning-Based Abnormal Behavior Detection for Elderly Healthcare Using Consumer Network Cameras. *IEEE Transactions on Consumer Electronics*, 70(1):2414–2422. – Cité page 54.

Zhang, Y.-L., Li, L., Zhou, J., Li, X., and Zhou, Z.-H. (2018). Anomaly Detection with Partially Observed Anomalies. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, pages 639–646, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee. – Cité page 32.

Zhang, Z. and Deng, X. (2021). Anomaly detection using improved deep SVDD model with data structure preservation. *Pattern Recognition Letters*, 148:1–6. – Cité page 41.

Zhang, Z., He, Q., Tong, H., Gou, J., and Li, X. (2016). Spatial-temporal traffic flow pattern identification and anomaly detection with dictionary-based compression theory in a large-scale urban network. *Transportation Research Part C: Emerging Technologies*, 71:284–302. – Cité page 53.

Zhao, M., Chen, J., and Li, Y. (2018). A Review of Anomaly Detection Techniques Based on Nearest Neighbor. In *2018 International Conference on Computer Modeling, Simulation and Algorithm (CMSA 2018)*, pages 290–292. Atlantis Press. – Cité page 30.

Zhao, Y. and Hryniewicki, M. K. (2018). XGBOD: Improving Supervised Outlier Detection with Unsupervised Representation Learning. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, Rio de Janeiro. IEEE. – Cité pages 26, 29, 31, 33, 39, 45, 94 et 103.

Zhao, Y., Nasrullah, Z., and Li, Z. (2019). PyOD: A python toolbox for scalable outlier detection. *Journal of Machine Learning Research*, 20(96):1–7. – Cité page 113.

Zhao, Y., Zheng, G., Mukherjee, S., McCann, R., and Awadallah, A. (2023). Admoe: Anomaly detection with mixture-of-experts from noisy labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 4937–4945. – Cité page 45.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. (2024). Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36. – Cité page 187.

Zheng, Z., Jeong, H.-Y., Huang, T., and Shu, J. (2017). KDE based outlier detection on distributed data streams in multimedia network. *Multimedia Tools and Applications*, 76(17):18027–18045. – Cité page 38.

Zhou, C. and Paffenroth, R. C. (2017). Anomaly Detection with Robust Deep Autoencoders. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 665–674, Halifax NS Canada. ACM. – Cité page 42.

Zhou, J., Kwan, C., Ayhan, B., and Eismann, M. T. (2016). A novel cluster kernel RX algorithm for anomaly and change detection using hyperspectral images. *IEEE Transactions on Geoscience and Remote Sensing*, 54(11):6497–6504. – Cité page 21.

Zhou, Y., Hu, W., Min, Y., Zheng, L., Liu, B., Yu, R., and Dong, Y. (2017). A semi-supervised anomaly detection method for wind farm power data preprocessing. In *2017 IEEE Power & Energy Society General Meeting*, pages 1–5. – Cité page 32.

Zhou, Y., Liu, P., and Qiu, X. (2022a). KNN-Contrastive Learning for Out-of-Domain Intent Classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5129–5141, Dublin, Ireland. Association for Computational Linguistics. – Cité page 39.

Zhou, Y., Song, X., Zhang, Y., Liu, F., Zhu, C., and Liu, L. (2022b). Feature Encoding with AutoEncoders for Weakly-supervised Anomaly Detection. *IEEE Transactions on Neural Networks and Learning Systems*, 33(6):2454–2465. – Cité page 36.

Zhou, Z.-H. (2012). *Ensemble Methods: Foundations and Algorithms*. CRC press. – Cité page 45.

Zhou, Z.-H. (2018). A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53. – Cité page 33.

Zhou, Z.-H. (2021). *Machine Learning*. Springer nature. – Cité page 27.

Zhou, Z.-H. and Zhou, Z.-H. (2021). *Ensemble Learning*. Springer. – Cité page 45.

Zhuang, H., Wang, C., Tao, F., Kaplan, L., and Han, J. (2017). Identifying Semantically Deviating Outlier Documents. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2748–2757, Copenhagen, Denmark. Association for Computational Linguistics. – Cité page 21.

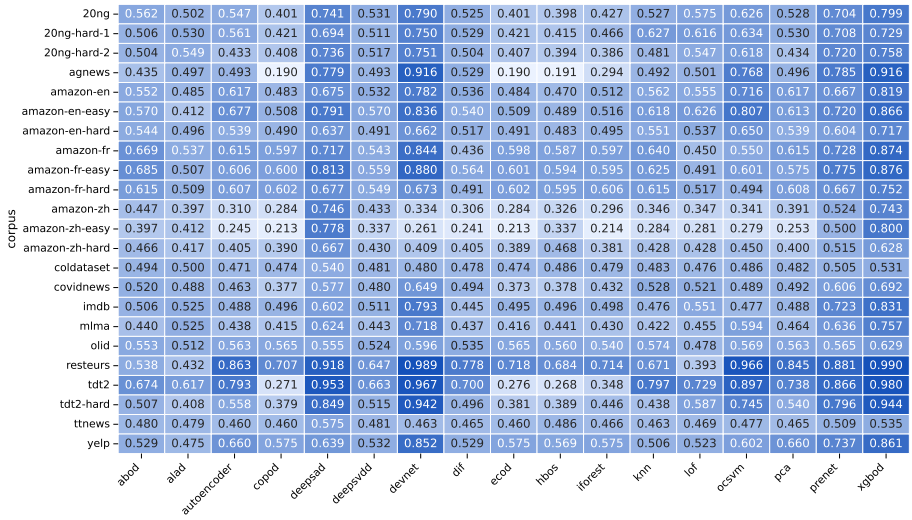
Zimek, A., Gaudet, M., Campello, R. J., and Sander, J. (2013). Subsampling for efficient and effective unsupervised outlier detection ensembles. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 428–436, New York, NY, USA. Association for Computing Machinery. – Cité page 45.

Zong, B., Song, Q., Min, M. R., Cheng, W., Lumezanu, C., Cho, D., and Chen, H. (2018). Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International Conference on Learning Representations*. – Cité page 39.

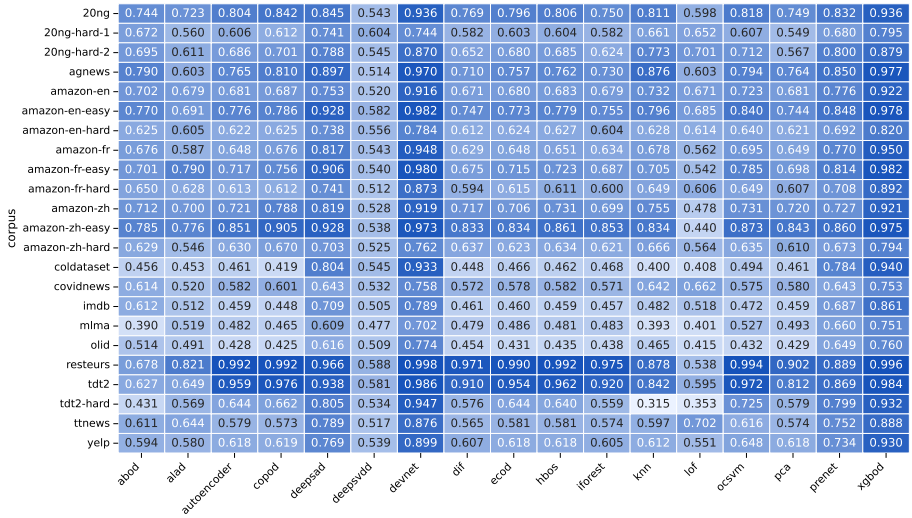
Zou, D., Xiang, Y., Zhou, T., Peng, Q., Dai, W., Hong, Z., Shi, Y., Wang, S., Yin, J., and Quan, H. (2023). Outlier detection and data filling based on KNN and LOF for power transformer operation data classification. *Energy Reports*, 9:698–711. – Cité page 40.

ANNEXE A

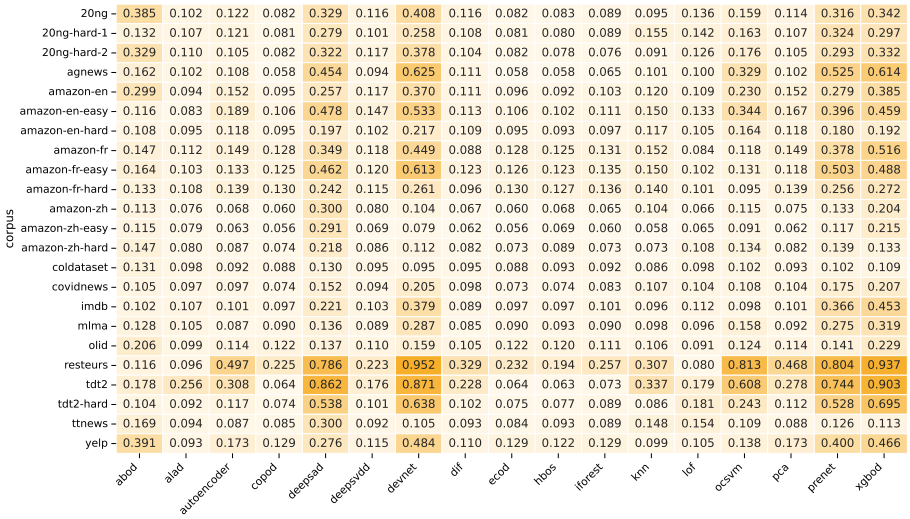
RÉSULTATS EXPÉRIMENTAUX : PARTIE II



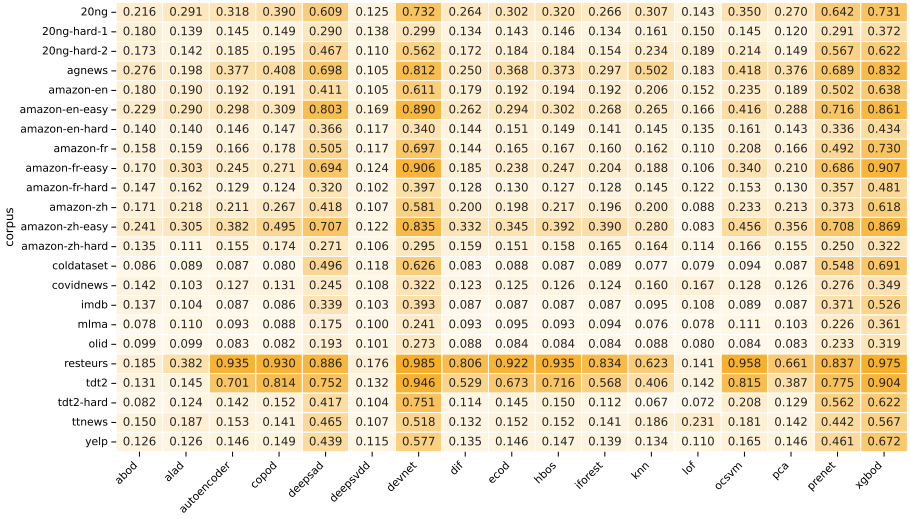
Moyenne de l'AUCROC sur 2 essais indépendants avec validation croisée à 5 plis pour 17 algorithmes appliqués à 17 corpus (23 sous-corpus) représentés par **TF-IDF**.



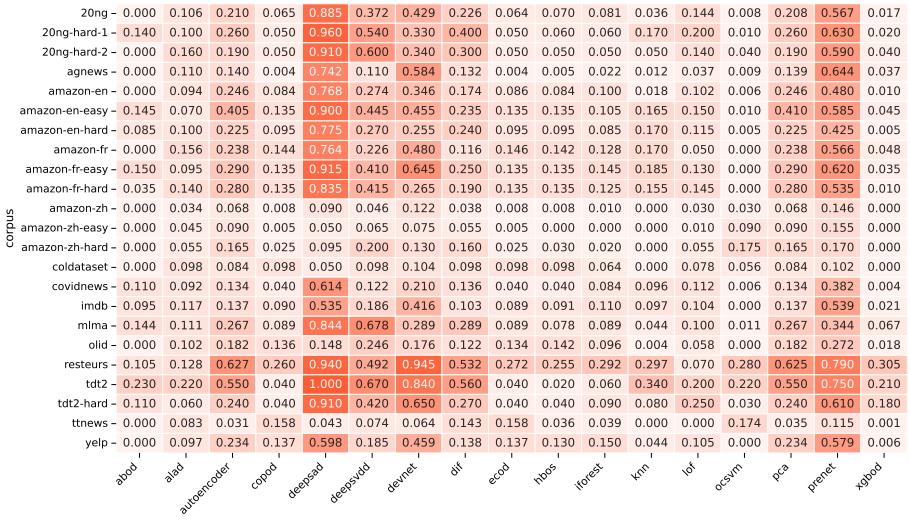
Moyenne de l'AUCROC sur 2 essais indépendants avec validation croisée à 5 plis pour 17 algorithmes appliqués à 17 corpus (23 sous-corpus) représentés par **SBERT-multi-mpnet**.



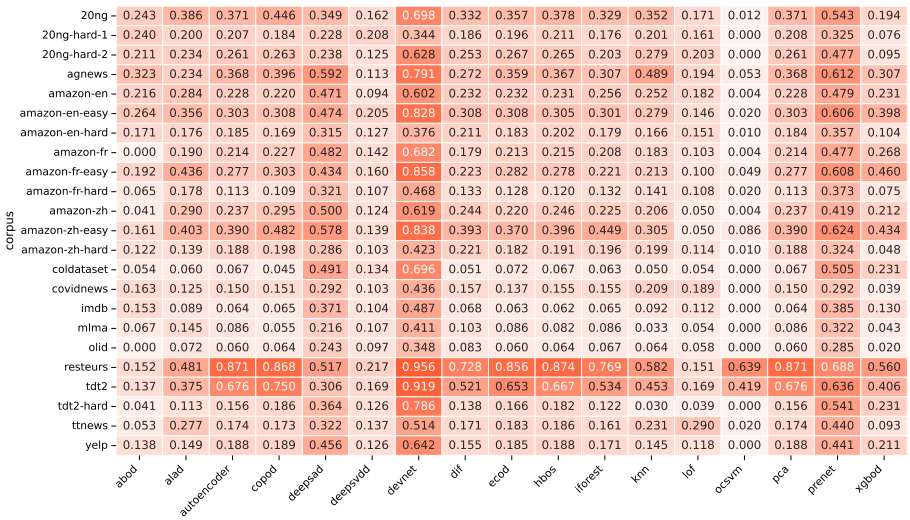
Moyenne de l'AUCPR sur 2 essais indépendants avec validation croisée à 5 plis pour 17 algorithmes appliqués à 17 corpus (23 sous-corpus) représentés par **TF-IDF**.



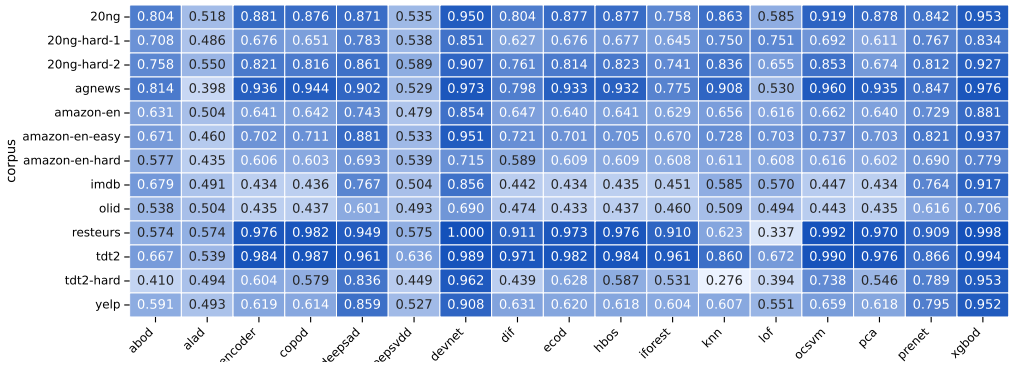
Moyenne de l'AUCPR sur 2 essais indépendants avec validation croisée à 5 plis pour 17 algorithmes appliqués à 17 corpus (23 sous-corpus) représentés par **SBERT-multi-mpnet**.



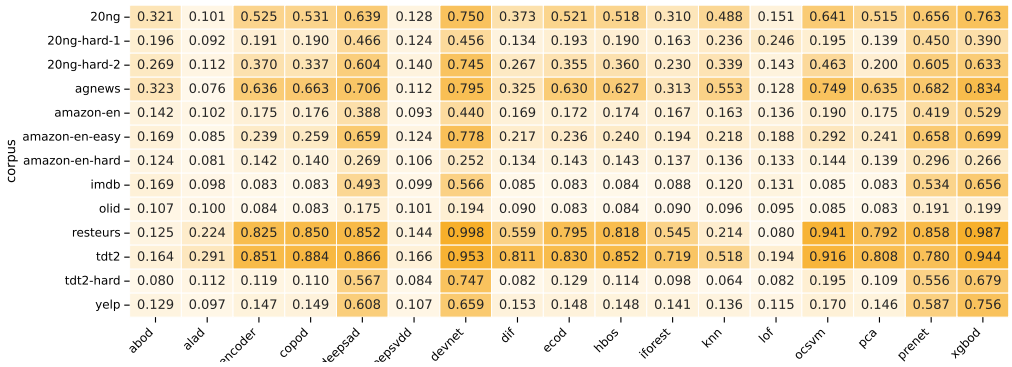
Moyenne du F-score sur 2 essais indépendants avec validation croisée à 5 plis pour 17 algorithmes appliqués à 17 corpus (23 sous-corpus) représentés par **TF-IDF**.



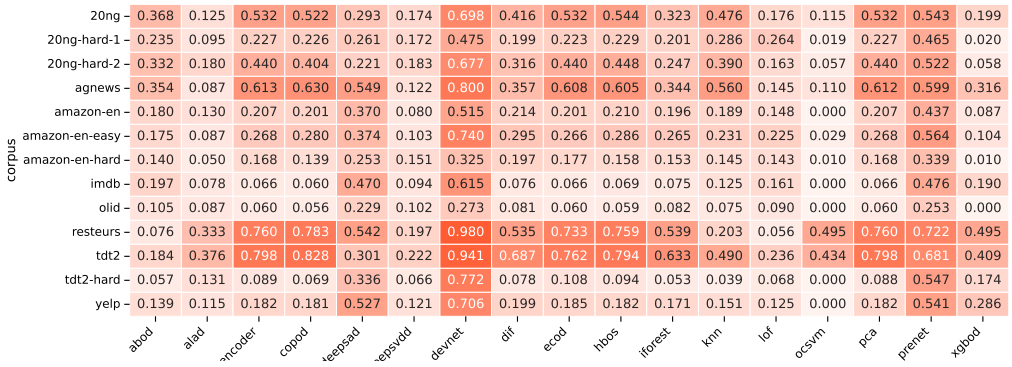
Moyenne du F-score sur 2 essais indépendants avec validation croisée à 5 plis pour 17 algorithmes appliqués à 17 corpus (23 sous-corpus) représentés par **SBERT-multi-mpnet**.



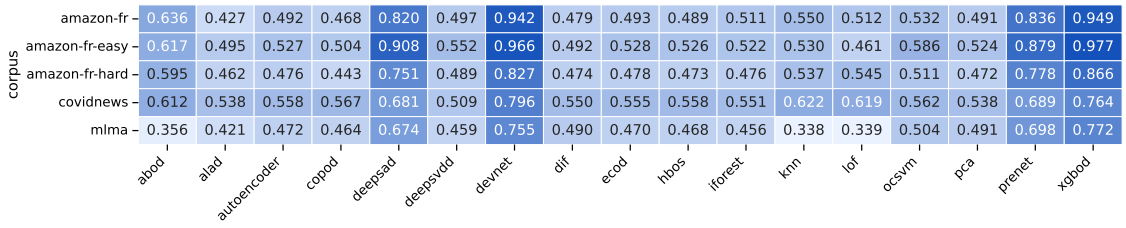
Moyenne de l'AUCROC sur 2 essais indépendants avec validation croisée à 5 plis pour 17 algorithmes appliqués à 13 (sous-)corpus anglais représentés par **SBERT-mpnet**.



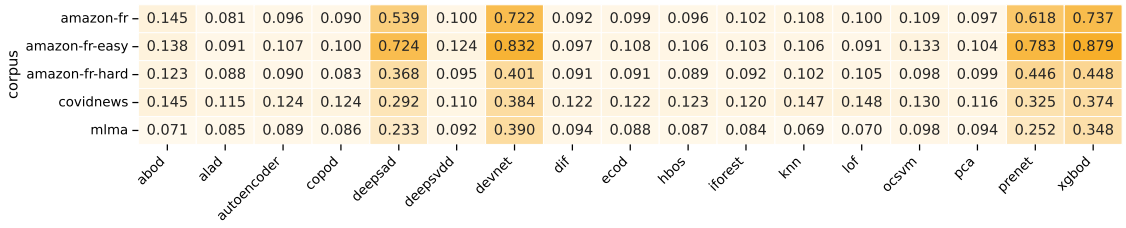
Moyenne de l'AUCPR sur 2 essais indépendants avec validation croisée à 5 plis pour 17 algorithmes appliqués à 13 (sous-)corpus anglais représentés par **SBERT-mpnet**.



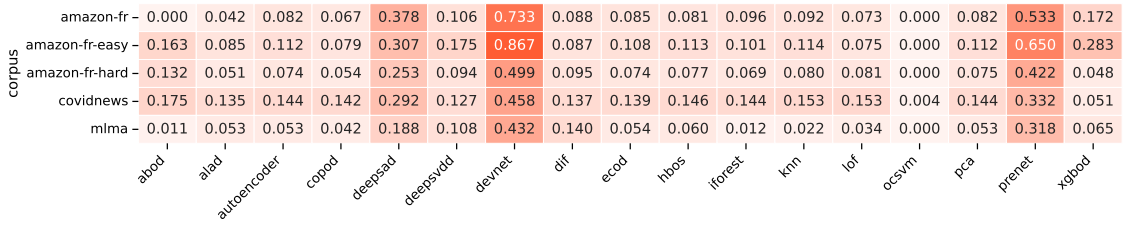
Moyenne du F-score sur 2 essais indépendants avec validation croisée à 5 plis pour 17 algorithmes appliqués à 13 (sous-)corpus anglais représentés par **SBERT-mpnet**.



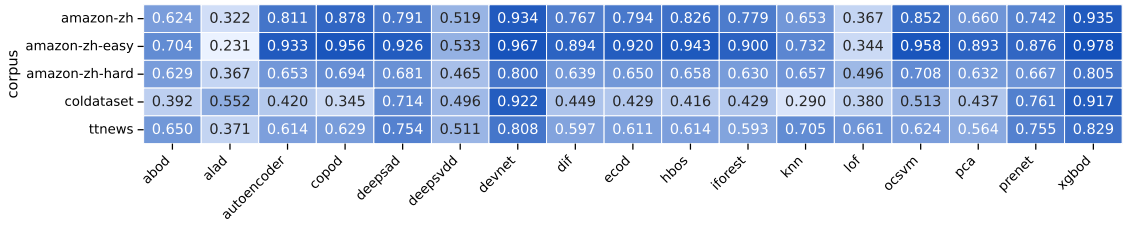
Moyenne de l'AUCROC sur 2 essais indépendants avec validation croisée à 5 plis pour 17 algorithmes appliqués à 5 (sous-)corpus français représentés par **SBERT-camembert-large**.



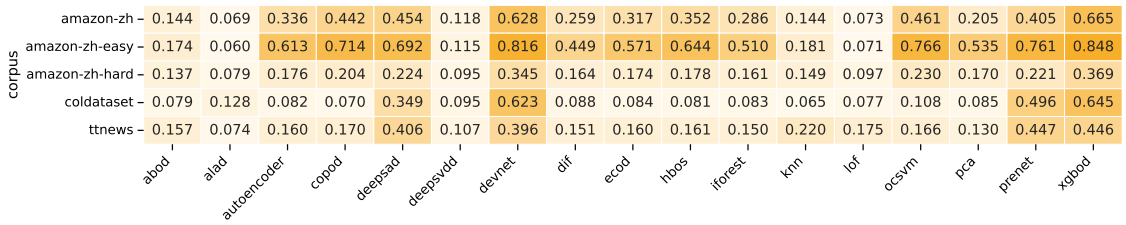
Moyenne de l'AUCPR sur 2 essais indépendants avec validation croisée à 5 plis pour 17 algorithmes appliqués à 5 (sous-)corpus français représentés par **SBERT-camembert-large**.



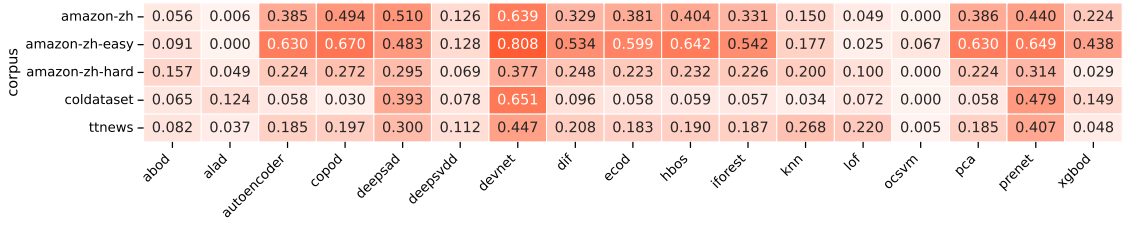
Moyenne du F-score sur 2 essais indépendants avec validation croisée à 5 plis pour 17 algorithmes appliqués à 5 (sous-)corpus français représentés par **SBERT-camembert-large**.



Moyenne de l'AUCROC sur 2 essais indépendants avec validation croisée à 5 plis pour 17 algorithmes appliqués à 5 (sous-)corpus chinois représentés par **SBERT-chinese-nli**.



Moyenne de l'AUCPR sur 2 essais indépendants avec validation croisée à 5 plis pour 17 algorithmes appliqués à 5 (sous-)corpus chinois représentés par **SBERT-chinese-nli**.



Moyenne du F-score sur 2 essais indépendants avec validation croisée à 5 plis pour 17 algorithmes appliqués à 5 (sous-)corpus chinois représentés par **SBERT-chinese-nli**.

ANNEXE B

GRANDS MODÈLES DE LANGUE

Le tableau récapitule diverses familles de grands modèles de langues (LLMs) abordées dans la thèse, avec des informations clés sur leur créateur, leur capacité de traitement, leur score moyen lors des évaluations, et leur accessibilité. Voici un détail des colonnes du tableau :

1. **Modèle** : Cette colonne indique le nom du modèle en question. Il peut s'agir de variantes de la même famille (par exemple, GPT-4 et GPT-4o).
2. **Créateur** : Le créateur du modèle, souvent une entreprise ou une organisation de recherche, qui a conçu, développé et publié le modèle.
3. **Contexte** : Représente la capacité maximale du modèle en termes de nombre de tokens pouvant être traités simultanément. Cela correspond à la taille de la fenêtre contextuelle, ce qui signifie le nombre de tokens que le modèle peut analyser ou générer en une seule fois.
4. **Score** : Il s'agit de la performance moyenne du modèle basée sur des évaluations spécifiques de l'intelligence artificielle. Les scores sont dérivés de tests publics tels que MMLU, GPQA, Math, et HumanEval.
5. **Accessibilité** : Indique si le modèle est accessible en open-source ou s'il est propriétaire, c'est-à-dire contrôlé par une entreprise et généralement accessible via une API ou une licence payante.

Modèle	Créateur	Contexte	Score	Accessibilité
GPT-3.5 Turbo	OpenAI	16k tokens	52	Propriétaire
GPT-4 Turbo	OpenAI	128k tokens	74	
GPT-4o	OpenAI	128k tokens	77	
Llama 3 8B	Meta	8k tokens	46	Open-source
Llama 3 70B	Meta	8k tokens	62	
Mistral 7B	MistralAI	33k tokens	24	Open-source
Mixtral 8x7B	MistralAI	33k tokens	42	
Mixtral 8x22B	MistralAI	65k tokens	61	
Gemini 1.0 Pro	Google	33k tokens		Propriétaire
Gemini 1.5 Flash	Google	1M tokens		
Gemini 1.5 Pro	Google	2M tokens		

ANNEXE C

GLOSSAIRE ET ABRÉVIATIONS

C.1 Glossaire

ajustement
Anglais : *fine-tuning*
Alias : réglage ; affinement
Pages : 69, 70, 271

apprentissage *few-shot*
Anglais : *few-shot learning*
Pages : 32, 175, 271

apprentissage en contexte
Anglais : *in-context learning*
Pages : 165, 271

apprentissage à base de prompt
Anglais : *prompt learning*
Pages : 165, 271

auto-attention
Anglais : *self-attention*
Pages : 162, 271

capacités émergentes
Anglais : *emergent abilities*
Pages : 164, 271

chaîne de pensée
Anglais : *chain of thought*
Acronyme : CoT
Pages : 166, 177, 271

estimation du maximum de vraisemblance
Anglais : *maximum likelihood estimation*
Acronyme : MLE
Pages : 38, 271

grands modèles de langue
Anglais : *large langue model(s)*

Acronyme : LLM(s)

Alias : grand modèle de langage ; grand modèle linguistique ; modèle de langage/langue de grande taille ; modèle de langage/langue à grande échelle

Pages : 72, 159, 271

ingénierie de prompt
Anglais : *prompt engineering*

Pages : 160, 165, 271

ingénierie des caractéristiques
Anglais : *feature engineering*

Pages : 157, 271

lois d'échelle
Anglais : *scaling laws*

Pages : 164, 271

modèles de langue pré-entraînés
Anglais : *pre-trained language model(s)*

Acronyme : PLM(s)

Pages : 58, 158, 271

modèles de langue statistiques
Anglais : *statistical language model(s)*

Acronyme : SLM(s)

Pages : 156, 271

modèles de langues neuronaux
Anglais : *neural language model(s)*

Acronyme : NLMs

Pages : 157, 271

modélisation de langue masquée
Anglais : *masked language modeling*

Acronyme : MLM

Pages : 71, 159, 271

traitement automatique des langues naturelles
Anglais : *natural language processing, NLP*

Acronyme : TALN

Alias : traitement automatique des langues, TAL

Pages : 57, 153, 155, 271

transformeur

Anglais : *transformer*

Alias : transformateur ; modèle auto-attentif

Pages : 162, 271

C.2 Abréviations

Abréviation	Forme Complète
ABOD	***A**ngle-**B**ased **O**utlier **D**etection*
ALAD	***A**dversarially **L**earned **A**nomaly **D**etection*
AUCROC	***A**rea **U**nder the **R**eceiver **O**perating **C**haracteristic **C**urve*
AUCPR	***A**rea **U**nder the **P**recision-**R**ecall **C**urve*
CDF	***C**umulative **D**istribution **F**unction*
COPOD	***COP**ula-based **O**utlier **D**etection*
CoT	***C**hain **o**f **T**hought*
CVDD	***C**ontext **V**ector **D**ata **D**escription*
DL	***D**eep **L**earning*
DM	***D**ata **M**ining*
DeepSAD	***Deep S**emi-supervised **A**nomaly **D**etection*
DeepSVDD	***Deep S**upport **V**ector **D**ata **D**escription*
DevNET	*Anomaly Detection with **Dev**iation **NET**works*
DIF	***D**eep **I**solation **F**orest*
ECDF	***E**mpirical **C**umulative **D**istribution **F**unction*
GAN	***G**enerative **A**dversarial **N**etwork*
GPT	***G**enerative **P**re-trained **T**ransformer*
GMM	***G**aussian **M**ixture **M**odel*
HBOS	***H**istogram-**B**ased **O**utlier **S**core*
KNN	***K**-**N**earest **N**eighbors*
LLM	***L**arge **L**anguage **M**odel*
LOF	***L**ocal **O**utlier **F**actor*
ML	***M**achine **L**earning*
OCC	***O**ne-**C**lass **C**lassification*
OCSM	***O**ne-**C**lass **S**upport **V**ector **M**achines*
PCA	***P**rincipal **C**omponent **A**nalysis*
PLM	***P**re-trained **L**anguage **M**odel*
PReNET	***P**airwise **Re**lation Prediction-based Ordinal Regression **NET**work*
PU	***P**ositive-**U**nlabeled (Semi-supervised Learning)*
NO	***N**ormal-**O**nly (Semi-supervised Learning)*
RLHF	***R**einforcement **L**earning from **H**uman **F**eedback*
MoE	***M**ixture **o**f **E**xperts*
SVDD	***S**upport **V**ector **D**ata **D**escription*
XGBOD	*EX**treme **G**radient **B**oosting **O**utlier **D**etection*

Institut national
des langues
et civilisations orientales

Yizhou XU

Détection d'anomalies dans les textes pour la veille

Résumé *: Le travail présenté dans cette thèse, mené au sein de Chapsvision dans le cadre d'une convention CIFRE, a pour objectif de développer et affiner des méthodes pour identifier des anomalies dans les textes, avec une attention particulière portée aux scénarios de veille stratégique. La détection d'anomalies, essentielle pour identifier erreurs, changements critiques ou activités suspectes, a été peu explorée dans le contexte des données textuelles.*

Pour combler cette lacune, notre approche repose sur deux axes. D'une part, des algorithmes d'apprentissage automatique conçus pour d'autres types de données (séries temporelles, données tabulaires, images) ont été adaptés pour traiter les spécificités des textes via des méthodes semi-supervisées, non supervisées et faiblement supervisées. D'autre part, nous avons intégré les modèles les grands modèles de langue (LLMs) dans notre solutionpour la détection d'anomalies, en expérimentant avec diverses stratégies de conception de prompts.

Les contributions de cette recherche permettent d'avancer l'état de l'art dans le domaine de la détection d'anomalies dans les données textuelles, offrant des perspectives prometteuses pour des applications en veille stratégique et au-delà.

Mots-clés *: détection d'anomalies, veille stratégique, apprentissage automatique, grands modèles de langue (LLMs)*

Anomaly Detection in Texts for Monitoring

Abstract *: The work presented in this thesis, conducted at Chapsvision under a CIFRE agreement, aims to develop and refine methods for identifying anomalies in texts, with a particular focus on strategic monitoring scenarios. Anomaly detection, crucial for identifying errors, critical changes, or suspicious activities, remains underexplored for text data.*

To address this gap, our approach is based on two axes. On the one hand, machine learning algorithms originally designed for other data types (e.g., time series, tabular data, images) were adapted for textual data using semi-supervised, unsupervised, and weakly supervised methods. On the other hand, we leveraged Large Language Models (LLMs) for anomaly detection, experimenting with various strategies for designing tailored prompts.

The contributions of this research advance the state of the art in the field of anomaly detection in textual data, offering promising prospects for strategic monitoring applications and beyond.

Keywords *: Anomaly Detection, Strategic Monitoring, Machine Learning, Large Language Models (LLMs)*