MASTER EN INGÉNIEUR DE GESTION À FINALITÉ SPÉCIALISÉE EN DATA SCIENCE

Evolution des usages du machine learning dans différents domaines d'applications

étude du reflet dans la littérature scientifique. (Analyse réalisée à l'aide de techniques de text mining.)

Stoffel, Annabelle

Award date:
2021

Awarding institution:
Universite de Namur

Evolution des usages du machine learning dans différents domaines d'applications : étude du reflet dans la littérature scientifique.

Analyse réalisée à l'aide de techniques de text mining.

Annabelle STOFFEL

Directeur: Prof. I. LINDEN

Mémoire présenté
en vue de l'obtention du titre de
Master 120 en ingénieur de gestion, à finalité spécialisée
en data science

ANNEE ACADEMIQUE 2020-2021

Université de Namur, ASBL
Faculté des Sciences économiques, sociales et de gestion – Département des Sciences de gestion

Rempart de la Vierge 8, B-5000 Namur, Belgique, Tel. +32 [0]81 72 49 58/48 41

Remerciements

Pour la réalisation de ce mémoire, j'ai reçu l'aide et le soutien de nombreuses personnes que je remercie.

Je remercie d'abord ma promotrice, Madame Linden, pour sa disponibilité, pour ses conseils et pour le temps investi à la relecture de ce mémoire.

Je remercie également les autres enseignants de l'Université de Namur, pour la chance que j'ai eue de participer à leurs cours et pour la connaissance qu'ils m'ont transmise afin que je puisse réaliser ce mémoire.

Enfin, je souhaite remercier mon entourage qui m'a soutenue et encouragée tout au long de la rédaction de ce mémoire.

Table des matières

Table des illustrations :

1 Introduction

Considéré par certain (L, 2020) comme « une innovation phare de ce début de XXème siècle », l'apprentissage automatique, mieux connu sous le nom machine learning (ML) fait régulièrement le titre de la presse. L'apprentissage automatique « aborde la question de savoir comment construire des programmes informatiques qui améliorent leurs performances dans une tâche donnée grâce à l'expérience » ((MITCHELL, 1997) cité dans (FRENAY, 2019). L'augmentation de son utilisation va de paire avec l'évolution du Big Data et la collecte de données de plus en plus nombreuses et utiles pour informer ou prendre une décision. L'augmentation des données ainsi que « l'apparition de processeurs capables d'effectuer d'importantes quantités de calculs et l'introduction d'algorithmes plus sophistiqués » (GOMAERE, 2019) ont permis d'avoir de plus en plus d'objets dotés d'intelligence artificielle. Ce sont par exemple des robots, des systèmes GPS ou encore des systèmes de recommandations des sites internet. L'intelligence artificielle est utilisée dans de nombreux domaines tels que « le domaine militaire, le secteur de la finance, la médecine, la robotique, les jeux vidéos, les transports ou les industries » (GROUPE MADEINFUTURA, 2020). Une majorité de ces nouveaux objets s'appuie sur des techniques de machine learning. Il est attendu que l'utilisation d'intelligence artificielle, et en particulier du machine learning, augmentera encore considérablement les années à venir, avec une valeur de ce marché mondial « passant de 2.4 milliards de dollars en 2017 à une valeur attendue de 59.8 milliards de dollars en 2025 » (DEPARTMENT STATISTA RESEARCH, 2019).

Les techniques de machine learning sont nombreuses. Citons notamment les arbres de décisions, les machines à vecteurs de support (SVMs en anglais), les réseaux de neurones, les régressions linéaires ou logistiques, le clustering, etc. Chacune de ces techniques est appropriée pour certains objectifs et a des avantages et des inconvénients. L'utilisation de ces techniques sera donc différente en fonction du domaine et du type d'application. Le nombre de techniques de machine learning, la multiplication et la diversification de leurs usages qui ne cessent de progresser, amènent à se poser la question suivante :

« Comment l'usage des techniques de machine learning a-t-il évolué au sein de différents domaines d'applications ?»

Les résultats de cette recherche fournissent une revue de l'utilisation des techniques de ML dans différents domaines d'applications et leurs évolutions. Ils permettent également de déterminer quelle technique est la plus probable d'être efficiente lors du lancement d'un

nouveau projet utilisant du machine learning dans un domaine particulier et quelles techniques ont tendance à devenir obsolètes.

Au vu de l'immensité du nombre d'applications utilisant des techniques de machine learning, il est impossible de lire un par un l'entièreté des articles concernant des applications du machine learning dans différents domaines. Dans ce mémoire, l'analyse est réalisée par l'application de techniques du text mining sur un ensemble d'articles publiés dans le journal *Expert Systems with Applications* entre 1990 et 2020. Cette analyse extrait de l'information telle que l'évolution des techniques de machine learning dans le temps, les techniques les plus utilisées de manière générale, etc. Tout cela de manière automatique sur base des différents articles scientifiques collectés. De plus, l'analyse via text mining va également nous permettre de déterminer les différents domaines d'applications qui utilisent du machine learning.

Afin de répondre à la question de recherche, ce mémoire débute par un état de l'art des différents travaux réalisés dans différents domaines identifiant les techniques de ML les plus utilisées sur une période de temps ou à un moment précis. Le second chapitre de ce mémoire présente la méthodologie. Celle-ci fournit un background méthodologique, l'explication du choix du journal et la présentation du processus. Les chapitres suivants décrivent les phases de notre processus à savoir : la préparation des données, la modélisation de sujet et l'analyse des résultats. Enfin, une conclusion est tirée afin de mettre en avant les principales contributions de ce travail.

2 Revue de la littérature

Une étude des travaux traitant de l'évolution des usages des techniques du machine learning dans certains domaines permettra de comparer les résultats obtenus avec ceux de ces travaux. Les analyses d'évolution étant limitées, ce travail reprend également des articles présentant une overview des utilisations à différentes périodes. Cela va nous permettre de faire le point sur l'usage à différents moments. Le choix de ces travaux s'est fait en fonction de certains domaines d'applications. Le choix des domaines analysés s'est basé sur les domaines d'applications de l'IA les plus courants à savoir : la santé, la finance, l'industrie, climat/énergie, la gestion d'entreprise et l'agriculture. Afin de couvrir toute la période d'analyse, des travaux à différentes dates ont été choisis. Les travaux ont été analysés en mettant l'accent sur l'objet d'étude ainsi que les conclusions. L'analyse de ces études est présentée par domaine d'applications. Le Tableau 2.1 fournit un récapitulatif des articles analysés.

2.1 Médecine

Dans le domaine médical, plusieurs travaux ont été réalisés. Nous allons analyser ici les travaux menés par KONONENKO (2001) et ZHANG, TAN, HAN & ZHU (2017). Le premier analyse le problème de diagnostic via les différentes techniques de classification suivantes : la classification bayésienne, les réseaux de neurones et les arbres de décisions. Ces techniques sont analysées sur différents critères : la performance, la transparence, l'explication, la réduction et la manipulation des données manquantes. Pour ce qui est du critère de performance, il y a peu de différence entre les différentes techniques. Etant donné qu'un point important en médecine est de savoir expliquer les résultats, l'étude prouve que ce sont les arbres de décisions et la classification bayésienne qui sont préférés aux réseaux de neurones. Les arbres de décisions sont mêmes préférés à la classification bayésienne car c'est la seule technique qui permet de répondre au critère de réduction du nombre de tests, c'est-à-dire qu'il a une facilité à « sélectionner un sous-ensemble approprié d'attributs pendant le processus» (KONONENKO, 2001). Il démontre également que pour améliorer la fiabilité et la compréhension de ces classifications, il est intéressant de combiner les résultats obtenus avec les différentes techniques de classification. Pour ce qui est du futur des années 2000, l'étude prédisait une lenteur dans l'acceptation de l'utilisation du ML pour diagnostiquer des maladies. Celle-ci était expliquée par des possibilités techniques restreintes et une complexité pour les médecins d'utiliser de nouveaux outils.

L'étude plus récente menée par ZHANG, TAN, HAN & ZHU (2017) montre l'évolution du machine learning et son évolution vers le deep learning dans le domaine médical mais cet article s'intéresse spécifiquement à la découverte de nouveaux médicaments. L'utilisation des techniques de base du ML dans la découverte de nouveaux médicaments est utilisée depuis plus d'une dizaine d'années et spécialement le QSAR[1] model. Le QSAR model comprend l'analyse linéaire discriminante (LDA), les SVMs, les arbres de décisions, les random forest, l'algorithme des k plus proches voisins (KNN) et les réseaux de neurones artificiels. Toutefois, ces dernières années, ce modèle est limité dans le nombre et la variété des données utilisées. C'est pourquoi, de nos jours, les concepteurs de médicaments se tournent de plus en plus vers le deep learning. Cette évolution est notamment due à une augmentation des données disponibles et la vitesse de calcul des ordinateurs. Néanmoins, le deep learning a certaines limitations. Tout d'abord, la performance du modèle dépend du montant des données disponibles. Cela peut être problématique dans le domaine médical étant donné qu'énormément de données pharmaceutiques sont privées. Une autre limite est l'interprétation des données. Comme l'indiquait déjà KONONENKO (2001), l'explication des résultats en médecine est un point important. Malgré cela, il est fort probable que le deep learning devienne d'une grande aide pour la découverte de médicaments dans le futur.

Retenons de ces travaux qu'au début des années 2000, ce sont les arbres de décisions et la classification bayésienne qui devraient être le plus utilisés pour des questions d'interprétation des données. Au fil du temps, d'autres méthodes ont été utilisées et ont fait leurs preuves comme les SVMs, les régressions ou réseaux de neurones. Cependant, ces dernières années ces méthodes ont tendance à être de moins en moins utilisées au profit du deep learning.

2.2 Finance

Dans le domaine de la finance, nous n'avons pas identifié de publication qui fournisse une vue sur l'ensemble des techniques de manière globale. C'est pourquoi nous distinguons ici 2 grandes activités de la finance : l'évaluation du crédit et la prédiction de faillite.

L'étude menée par SADATRASOUL, GHOLAMIAN, SIAMI & HAJIMOHAMMADI (2013) analyse différentes techniques de ML utilisées de 2000 à 2012 pour évaluer le risque de crédit pour un individu mais également pour des entreprises. Les résultats démontrent que les techniques les plus utilisées sont les réseaux de neurones, les

[1] « Modèles mathématiques qui peuvent être utilisés pour prédire les propriétés physicochimiques, biologiques et le devenir environnemental des composés à partir de la connaissance de leur structure chimique » (EUROPEAN CHEMICALS AGENCY (ECHA))

méthodes d'ensemble et des SVMs. Ils soulignent également une forte utilisation des méthodes utilisant des règles comme les arbres de décisions ou les classifieurs basés sur des règles. Cela s'explique par le fait qu'ils ont un pouvoir explicatif des résultats obtenus, ce qui permet à ces modèles d'être mieux compris par les citoyens ou entreprises demandant le crédit. Des résultats semblables sont obtenus dans l'étude de DASTILE, CELIK & POTSANE (2020) avec comme méthodes les plus utilisées les SVMs et les réseaux de neurones artificiels. Les réseaux de neurones profonds démontrent encore peu d'utilisation liée au fait de leur récente apparition. Cependant, l'étude met en évidence que les réseaux de neurones convolutionnels (CNN) ont des performances de classification plus précises que les méthodes d'ensembles. Contrairement à l'étude de SADATRASOUL, GHOLAMIAN, SIAMI & HAJIMOHAMMADI (2013), l'analyse montre que la grande majorité des techniques utilisées (92%) n'ont pas de pouvoir explicatif des données alors que ces explications sont une nécessité dans l'évaluation du risque de crédit.

En ce qui concerne la prédiction de faillite, les résultats de l'étude menée sur des données de 1968 à 2017 par QU, QUAN, LEI & SHI (2019) démontrent que ce sont les réseaux de neurones, les SVMs et les arbres de décisions les plus utilisés. Dans cette analyse, les modèles statistiques et de machine learning ont été distingués. Dans les méthodes statistiques, c'est la régression logistique qui est la plus utilisée. Celle-ci est même plus utilisée que les réseaux de neurones. BOSE & MAHAPATRA (2001) rajoutent également que les modèles basés sur des règles d'inférence et le raisonnement basé sur des cas sont utiles pour les prédictions dans la finance.

L'étude menée par LIN, HU & TSAI (2011) sur la prédiction de faillite et le risque de crédit présente une étude qui analyse 130 journaux allant de 1995 à 2010. Les résultats rejoignent les résultats démontrés dans les autres études analysées, c'est-à-dire que les réseaux de neurones et SVMs sont fortement utilisés. Cependant, ils complètent le top 3 des techniques de classification simples utilisées avec l'algorithme génétique apparaissant à partir de l'année 2002. Les SVMs quant à eux apparaissent à partir de 2004 et les réseaux de neurones sont utilisés sur toute la période. Cet article met également l'accent sur les techniques de classification « soft » à savoir les ensembles[2] et les classifieurs hybrides[3], qui sont des techniques de classification plus utilisées que les simples classifieurs dans le domaine

[2] « combiner plusieurs techniques de classifications avec une probabilité pour chaque modèle » (LIN, HU, & TSAI, 2011)
[3] « combiner 2 ou plus de techniques de machine learning hétérogènes » (LIN, HU, & TSAI, 2011)

de la finance. Les classifications hybrides sont les plus utilisées avec notamment « SVMs et réseaux de neurones qui sont les plus utilisés pour la classification et l'algorithme génétique qui est largement utilisé pour optimiser les paramètres pour entrainer ces 2 méthodes de classification » (LIN, HU, & TSAI, 2011).

Au vu de ces travaux réalisés dans différentes activités de la finance, nous pouvons dire que les réseaux de neurones sont une technique à forte utilisation depuis toujours dans ce domaine. L'apparition en 2002 des algorithmes génétiques a également eu un impact dans les techniques utilisées. Ensuite, vers 2004, les SVMs font leurs apparitions pour devenir une des techniques les plus utilisées. Les arbres de décisions restent une méthode à utilisation constante. Pour ce qui est de ces dernières années, les applications de la finance ont tendance à utiliser de plus en plus de réseaux de neurones profonds au vu de leurs performances.

2.3 Industrie

Pour l'analyse de la production en entreprise, SHARP, AK & HEDBERG JR (2018) analysent quels algorithmes de machine learning sont les plus utilisés et à quel niveau de la production de 2005 à 2017. Les résultats montrent que ce sont les réseaux de neurones et les SVMs qui sont les plus présents dans les étapes de production.

WUEST, WEIMER, IRGENS & THOBEN (2016) fournissent une vue d'ensemble sur toutes les applications qui ont été faites dans la manufacture à l'aide de technique de machine learning. L'étude a constaté qu'il y avait de plus en plus d'utilisations de SVMs malgré que ce soit un concept récent, dû à une haute performance de ce modèle. Ils démontrent aussi qu'il y a plus d'utilisation de techniques de ML supervisées que de non supervisées. Cela est notamment dû au fait que les entreprises ont une capacité à fournir des données labellisées.

GE, SONG, DING & HUANG (2017) analysent le rôle du data mining et de l'analyse dans les processus d'industrie entre 2000 et 2015. Ils découpent l'analyse en 4 catégories : apprentissage supervisé, non supervisé, par renforcement et semi supervisé. Le travail met en évidence que 80 à 90% des applications dans l'industrie se faisaient via des méthodes d'apprentissage supervisé et non supervisé. La méthode non supervisée la plus utilisée est l'analyse en composante principale (PCA) avec plus de 51% d'applications. Il y a ensuite la méthode d'apprentissage multiple, l'analyse en composantes indépendantes et la carte auto-organisée (type de réseau de neurones). Les auteurs ont également analysé les utilisations par types d'activités très présentes dans le domaine de l'industrie, c'est-à-dire pour la réduction de dimensionnalité, la détection de valeur aberrantes, le contrôle du processus et la visualisation

des données. La méthode d'apprentissage multiple est particulièrement utilisée pour la réduction de dimensionnalité et la visualisation des données. L'analyse des composantes est quant à elle fortement sollicitée dans le cas du contrôle du processus. L'utilisation de la carte auto-organisée se fait surtout dans la visualisation des données. La méthode k-means est très utilisée pour détecter les valeurs aberrantes. Si on regarde maintenant les analyses faites sur l'apprentissage supervisé, nous observons que les méthodes les plus utilisées sont les réseaux de neurones (29%), la régression des moindres carrés (28%) et les SVMs (16%). Il y a la régression à composantes principales et les arbres de décisions/random forest avec 9% d'utilisations. Les réseaux de neurones représentent 28% des utilisations pour le contrôle du processus, 34% pour la classification de défaut, 22% pour l'analyse de capteur et 28% pour la prédiction de qualité. La régression des moindres carrés représente quant à elle 28% également pour le contrôle de processus, 35% pour l'analyse de capteurs et 34% pour la prédiction de qualité. Elle a peu d'utilisation pour la classification de défauts. Par contre, les SVMs représentent plus de 39% des utilisations pour la classification de défauts.

Nous pouvons donc en déduire que les techniques les plus utilisées dans l'industrie sont les réseaux de neurones et la régression linéaire. A travers ces 3 articles, nous pouvons voir l'importance qu'ont pris les SVMs. En effet, WUEST, WEIMER, IRGENS & THOBEN (2016) indiquent que c'est une méthode récente qui prend de plus en plus d'ampleur. Ensuite, GE, SONG, DING & HUANG (2017) confirment cette forte utilisation en se hissant dans le top 3 des techniques supervisées. Enfin, SHARP, AK & HEDBERG JR (2018) montrent que les SVMs sont une des techniques les plus importantes.

2.4 Climat et énergie

Comme dans le secteur de la finance, nous n'avons pas identifié d'article sur le climat et l'énergie d'un point de vue global. Nous présentons ici une synthèse des travaux sur la prédiction solaire et les inondations.

L'objectif de VOYANT & al. (2017) est de « donner un aperçu des méthodes de prévisions de l'irradiation solaire utilisant des approches d'apprentissage automatique ». Une analyse de l'évolution de l'utilisation des réseaux de neurones et des SVMs est réalisée sur base de données provenant de cinq journaux parlant de l'énergie solaire entre 2000 et 2014. L'analyse montre que les réseaux de neurones sont plus utilisés que les SVMs. Le nombre d'apparition des réseaux de neurones augmente fortement à partir de 2012. Les SVMs quant à eux apparaissent vers l'année 2009. Ils connaissent une augmentation également à partir de 2012

mais celle-ci est inférieure à l'augmentation que subissent les réseaux de neurones. L'utilisation des réseaux de neurones est justifiée par sa précision dans les prévisions. Les auteurs indiquent que malgré que les techniques comme les arbres de régression, les random forest ou le boosting ne sont pas beaucoup utilisées, elles donnent également de bons résultats. L'article met également en avant qu'un ensemble de prédicteurs pourrait être utilisé afin d'obtenir de meilleurs résultats. Pour finir, les auteurs émettent des prédictions sur le futur : « les trois méthodes qui devraient généralement être utilisées dans les années à venir seront les SVMs, les arbres de régression et les random forests, car les résultats donnés sont très prometteurs et certaines études intéressantes seront certainement produites dans les prochaines années » (VOYANT, et al., 2017).

MOSAVI, OZTURK & CHAU (2018) se focalisent sur les méthodes de ML utilisées dans la prédiction d'inondations. L'analyse montre que ce sont les réseaux de neurones, les SVMs, les multilayer perceptrons, les arbres de décisions/random forest et système d'inférence neuro-flou, les wavelet neural networks et les systèmes d'ensemble de prédiction qui sont les plus utilisés. Ces techniques peuvent être divisées en techniques simples et techniques hybrides, c'est-à-dire les trois dernières techniques mentionnées. Les SVMs subissent une forte augmentation de 2011 à 2016. Les arbres de décisions ont également une tendance à être plus utilisés ces dernières années. Les réseaux de neurones connaissent quant à eux une grande évolution dans le nombre d'usage. L'étude montre également un intérêt de plus en plus fort pour les méthodes hybrides qui permettent d'obtenir de meilleures prédictions.

Ces études montrent que les réseaux de neurones sont utilisés depuis les années 2000. Depuis l'apparition des SVMs vers les années 2009, ceux-ci deviennent une méthode importante dans ce secteur. Les arbres de décisions sont également fortement utilisés. Les réseaux de neurones ont subi une forte augmentation d'utilisation notamment grâce à l'apparition du deep learning.

2.5 Gestion entreprise

NGAI, XIU & CHAU (2009) analysent des données de 2000 à 2006 sur base de 24 revues concernant la gestion de la relation client. Une des analyses porte sur le nombre d'applications des techniques de machine learning dans l'ensemble de ces données. Nous pouvons observer que ce sont dans l'ordre : les réseaux de neurones, les arbres de décision, les règles d'association, les régressions et l'algorithme génétique qui composent le top cinq des techniques les plus utilisées. L'apparition des réseaux de neurones dans le haut du classement est justifiée par le fait que ceux-ci peuvent être une technique de classification, de

regroupement mais également de prédiction. Les arbres de décisions et les techniques d'association permettent une meilleure interprétation des données, c'est pourquoi elles sont également dans le haut du classement.

Un autre point important dans la gestion d'entreprise concerne la prédiction de la demande dans la chaine d'offre. CARBONNEAU, LAFRAMBOISE & VAHIDOV (2008) effectuent une analyse sur cette activité. Les résultats montrent que les réseaux de neurones et les SVMs donnent de bonnes performances de prédiction. La technique la plus simple de régression donne également de bons résultats. L'analyse bayésienne a quant à elle de mauvaises performances.

SYAM & SHARMA (2018) réalisent une étude concernant l'impact que le ML et l'intelligence artificielle auront dans la gestion des ventes. Dans son étude, ils mettent en avant 3 techniques de ML qui sont très utilisées dans la gestion des ventes : les réseaux de neurones, les SVMs et le traitement du langage naturel. Cette dernière est justifiée par le fait que « au cours de la dernière décennie, nous avons également constaté une augmentation rapide de l'utilisation des moyens mobiles et basés sur le Web grâce auxquels l'organisation de vente peut contacter les clients » (SYAM & SHARMA, 2018).

Via ces trois articles, nous pouvons voir l'importance des réseaux de neurones et la régression dans la gestion d'entreprise. Il y a également une forte utilisation des arbres de décisions dans la gestion de la relation client. De plus, nous pouvons voir que les SVMs n'étaient pas mentionnés dans l'étude de NGAI, XIU & CHAU (2009) sur des données avant 2006, alors que dans les autres recherches effectuées plus tard, ceux-ci sont mentionnés. Cela explique que les SVMs étaient encore peu présents avant 2006.

2.6 Agriculture

LIAKOS & al. (2018) réalisent une étude des techniques de machine learning les plus employées de 2004 à 2018 dans différents sous-domaines de l'agriculture tels que la gestion de l'eau ou encore la production animale. Les techniques qui dominent sont respectivement les réseaux de neurones et les SVMs. Les réseaux de neurones sont particulièrement utilisés dans la détection de maladie et dans la gestion de l'eau et des sols. Les SVMs semblent avoir un taux d'usage semblable dans la détection de maladie, la prédiction du rendement, la qualité des récoltes et la production du bétail.

Domaine	Articles	Méthodes (classement)	Informations supplémentaires
Santé	(KONONENKO, 2001)	Arbre de décision Classification bayésienne Réseaux de neurones	Importance pour l'interprétation des données Combiner techniques
	(ZHANG, TAN, HAN, & ZHU, 2017)	LDA, svm, arbres de décisions, random forest, knn et réseaux de neurones artificiels. Puis **deep learning**	Limite deep learning : quantité de données et interprétation des données
Finance	(SADASTRASOUL, GHOLAMIAN, SIAMIi, & HAJIMOHAMMADI, 2013) : 2000 à 2012, risque de crédit	Réseaux de neurones, ensembles et svm Arbre de décision, classification basée sur les règles	Pouvoir explicatif des arbres de décisions et modèles basés sur les règles
	(DASTILE, CELIK, & POTSANE, 2020): risque de crédit	Svm Réseaux de neurones (CNN)	Cnn ont des performances de classification plus précises que les méthodes d'ensembles
	(QU, QUAN, LEI, & SHI, 2019) 1968 -2019	Réseaux de neurones Svm Arbre de décisions	Statistiques : régression logistique > réseaux de neurones
	(BOSE & MAHAPATRA, 2001)	Règles d'inférence et raisonnement basés sur cas	
	(LIN, HU, & TSAI, 2011) : 1995-2010	Réseaux de neurones Svm (à partir de 2004) Algorithme génétique (à partir de 2002)	Ensembles et classifieurs hybrides sont plus utilisés que les simples classifieurs.
Industrie	(SHARP, AK, & HEDBERG JR, 2018) : 2005 – 2017, étapes de productions	Réseaux de neurones Vvm	
	(WUEST, WEIMER, IRGENS, & THOBEN, 2016) : manufacture	De plus en plus de svm Utilisation supervisée > utilisation non supervisée	Beaucoup de données labellisées
	(GE, SONG, DING, & HUANG, 2017): 2000-2015	NS : PCA,Apprentissage multiple, Analyse composantes indépendantes,Carte auto-organisée S : Réseaux de neurones,Régression,svm, arbres de décision / random forest	Supervisé + non supervisé = 80-90% des applications Répartition des utilisations de techniques de ML par activités dans ce domaine
Climat / Energie	(VOYANT, et al., 2017) : prévision solaire, 2000-2014	Réseaux de neurones (à partir de 2012 forte augmentation) Svm (à partir de 2009 et forte augmentation en 2012)	Arbres de régression, les random forest ou le boosting donnent également de bons résultats Ensemble pour avoir meilleurs résultats Pour le futur : svm, arbre de régression et random forest
	(MOSAVI, OZTURK, & CHAU, 2018): inondations, 2011-2016	Réseaux de neurones (grande évolution) Svm (forte augmentation de 2011 à 2016 Multilayer perceptron Arbres de décisions/random forest (plus utilisés ces dernières années) Système d'inférence neuro-flou Wavelet neural networks Ensembles	Division en technique simple et hybride. Intérêt de plus en plus fort pour les méthodes hybrides car meilleures prédictions
Gestion entreprise	(NGAI, XIU, & CHAU, 2009) : gestion de la relation client, 2000-2006	Réseaux de neurones Arbres de décisions Règles d'association Régression Algorithme génétique	arbres de décisions et les techniques d'association permettent une meilleure interprétation des données
	(CARBONNEAU, LAFRAMBOISE, & VAHIDOV, 2008) : prédiction de la demande	Réseaux de neurones SVM Régression	Bayes mauvais Réseaux de neurones et SVM un peu meilleur que régression
	(SYAM & SHARMA, 2018) : Gestion des ventes	Réseaux de neurones SVM NLP	NLP car augmentation du contenu sur le web
Agriculture	(LIAKOS, BUSATO, MOSHOU, PEARSON, & BOCHTIS, 2018): 2004-2018	Réseaux de neurones Svm	Réseaux de neurones sont particulièrement utilisés dans la détection de maladie et dans la gestion de l'eau et des sols. Les svm semblent avoir un taux d'usage semblable dans la détection de maladie, la prédiction du rendement, la qualité des récoltes et la production du bétail.

Tableau 2.1: Revue de la littérature sur l'utilisation et l'évolution des techniques de machine learning (Source: Auteurs)

Le Tableau 2.1 résume les observations des articles, il nous permet de voir que les méthodes qui sont les plus utilisées au sein des différents domaines d'applications analysés sont les réseaux de neurones et les SVMs. L'apparition du deep learning vers les années 2010 a eu un effet sur les méthodes les plus utilisées, étant donné que nous voyons que cette méthode est de plus en plus choisie. Les arbres de décisions/random forest ne sont pas la méthode la plus utilisée mais interviennent quand même dans tous les domaines dû notamment à la grande capacité d'interprétation. Une autre observation est que plusieurs auteurs mettent également l'accent sur le fait que les méthodes d'ensembles devraient être plus utilisées afin d'obtenir de meilleurs résultats. Nous pouvons également voir que peu d'articles mettent l'accent sur l'évolution de l'usage des techniques dans le temps. C'est sur ce point que ce mémoire va pouvoir apporter sa contribution.

3 Méthodologie

Comme indiqué dans l'introduction, l'analyse se réalise par l'application de techniques du text mining sur un ensemble d'articles publiés dans le journal *Expert Systems with Applications*. Afin d'obtenir de bons résultats, il est nécessaire de suivre un processus bien précis. Ce chapitre comprend un background méthodologique, l'explication du choix du journal et le processus suivi afin de répondre à la question de recherche.

3.1 Background méthodologique

Avant d'aborder la méthodologie spécifique de ce mémoire, dressons un état de l'art des techniques du text-mining pour définir les différentes techniques existantes ainsi que la procédure à suivre pour l'application des techniques de text mining. Ensuite, un état des lieux sur quelques travaux réalisés à l'aide du text-mining sur des articles scientifiques fournit des exemples de cas et permet ainsi de s'inspirer ou comparer les procédures.

3.1.1 Les techniques de text mining

Le text mining est défini comme « la découverte par les ordinateurs de nouvelles informations inconnues, en extrayant automatiquement les informations de différentes ressources écrites » (GUPTA & LEHAL, 2009). Le text mining est également connu sous le nom « text data mining and knowledge discovery from textual databases » (DANG & AHMAD, 2014). Il existe énormément de techniques de text mining présentées dans différentes text books tels que celui de BENGFORT, BILBRO, & OJEDA (2018). Chaque technique se distingue par ses objectifs et ses avantages et inconvénients. Les techniques les plus courantes sont le résumé, la catégorisation/classification, le regroupement (clustering), la modélisation de sujet (topic modeling), l'extraction d'information, la visualisation de l'information, la réponse aux questions, le suivi de thème et l'analyse des sentiments. KOBAYASHI & al. (2018) fournissent un tableau intéressant mettant en lien la question qu'un utilisateur se pose quand il fait face à un texte et la technique de machine learning qui y répond et donnant des exemples d'applications dans la littérature. KAUSHIK & NAITHANI (2016) fournissent un bon résumé de ces différentes techniques et donnent également des outils de text mining disponibles sur le web et différents cas d'utilisation courants. Au vu du nombre de techniques disponibles, il semble intéressant de faire un récapitulatif de celles-ci afin de déterminer lesquelles seront susceptibles de nous aider dans notre recherche.

La première technique est le **résumé**. Cette technique permet de réduire la taille d'un texte en conservant les éléments importants du texte et son sens. Le résumé permet donc un

énorme gain de temps pour les utilisateurs qui ne sont plus obligés de lire l'entièreté du document pour déterminer si le document répond à leurs besoins. Les mots / phrases à garder ou à éliminer sont déterminés à partir d'un seuil de fréquence déterminé par l'auteur. Les techniques de ML les plus utilisées pour effectuer le résumé de texte sont : les réseaux de neurones, les arbres de décisions, les graphes sémantiques, les modèles de régression, la logique flou.

Une deuxième technique est la **catégorisation/classification**. Celle-ci vise à déterminer le sujet/ la catégorie d'un document sur base d'un modèle pré-entrainé avec des données labellisées. Etant donné que les catégories des données d'entrainement sont prédéfinies, différentes techniques de machine learning supervisées peuvent être utilisées pour réaliser la classification : la classification naïve bayésienne, les k plus proches voisins, les arbres de décisions, les SVMs, le boosting et les réseaux de neurones. De nos jours, il existe énormément d'applications de cette technique mais « la plus connue est certainement le filtrage de spam » (BENGFORT, BILBRO, & OJEDA, 2018).

La troisième technique est le **regroupement** (clustering). Cette technique vise à regrouper des documents considérés comme similaires. Il se peut que des groupes se superposent sur certains points ou soient complètement différents les uns des autres. La similarité des documents peut être évaluée par de multiples mesures : Jaccard, TF-IDF ou la similarité du cosinus. La première se mesure en indiquant le total des mots présents dans les deux documents (doc A et doc B) divisé par le nombre total de mots apparaissant dans l'ensemble de ces deux documents (doc A ou doc B). La mesure TF-IDF permet de mesurer la distance entre deux vecteurs en calculant le total des mots présents dans les deux documents (doc A et doc B) divisé par le nombre total de mots apparaissant dans tous le corpus. Enfin, la dernière mesure est également une mesure de vecteurs qui permet d'évaluer si les deux vecteurs sont dans la même orientation et donc similaires. Le clustering est différent de la catégorisation car cette dernière se base sur des sujets prédéfinis alors que le clustering ne demande pas de sujets prédéfinis. Il regroupe les documents similaires directement à partir des documents analysés. Etant donné que ce sont des données non labellisées, cette technique requiert l'utilisation de méthodes d'apprentissage non supervisé comme par exemple la méthode k-means ou alors le clustering hiérarchique, le biclustering[4] ou la factorisation par matrice non négative, le fuzzy clustering.

[4] Regroupement à la fois sur les lignes et les colonnes d'une matrice

La quatrième technique est la **modélisation du sujet** (topic modeling). Cette technique permet d'extraire des sujets abstraits dans l'ensemble du corpus. Pour réaliser la modélisation de sujet, trois méthodes sont connues : Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA) et Non-Negative Matrix factorisation (NNMF). LDA est un modèle probabiliste permettant d'obtenir les mots les plus représentatifs des sujets abstraits dans le corpus, et d'assigner à chaque document un ou plusieurs sujets. Par exemple, un document peut être associé à 30% au sport et à 70% à la nourriture. Et le sujet sport, peut être représenté par les mots « football, tennis, raquette, ballon ». LSA permet de trouver des sujets sur base de mots qui sont sémantiquement similaires. « LSA est parfois considéré comme meilleur pour apprendre des sujets descriptifs, qui est utile avec des longs documents et des corpus plus diffus » (BENGFORT, BILBRO, & OJEDA, 2018). La dernière permet de décomposer la matrice TF-IDF en 2 matrices : sujets et documents, de sorte que chaque valeur soit nulle ou positive.

La cinquième technique est **l'extraction d'information** permettant d'identifier des phrases importantes et relations entre les mots. Cela permettra d'obtenir des entités nommées comme les noms de personnes ou organisations, des lieux et des dates. L'extraction d'information est particulièrement utile pour obtenir des données structurées sur base d'un corpus de texte non structuré. Ce qui permet par la suite d'appliquer des techniques de data mining étant donné que les informations retrouvées peuvent être stockées dans une base de données structurées.

La sixième technique est la **visualisation de l'information**. Celle-ci permet de créer des cartes et autres objets visuels permettant ainsi à l'utilisateur de trouver rapidement de l'information et des liens entres différents concepts ou document provenant de longs documents textuels. « Le texte visualisé est facile à comprendre par rapport aux données numériques, car tout le monde peut facilement analyser les tendances à partir d'une image bien dessinée » (KAUSHIk & NAITHANI, 2016). Les différents outils de visualisation permettre d'obtenir de l'information des mots de plusieurs manières. Par exemple, « de voir l'évolution de l'utilisation d'un mot en fonction du temps, visualiser les relations entre les mots, visualiser la cooccurrence entre les mots, mettre certains mots en couleur en fonction de la place dans la phrase, visualiser la fréquence des mots » (BENGFORT, BILBRO, & OJEDA, 2018). La visualisation d'information est couramment utilisée dans l'analyse des réseaux sociaux (KAUSHIk & NAITHANI, 2016).

La septième technique est la **réponse aux questions** analysant comment donner la meilleure réponse à une question. Un exemple bien connu est l'utilisation de Google. « L'utilisateur demande à Google des questions ou recherche quelque chose, Google donne la meilleure réponse ou le lien pour cette question en cherchant les mots-clés de la question dans la base de données » (KAUSHIk & NAITHANI, 2016). Cette technique est utilisée avec d'autres techniques de text mining telles que « l'extraction d'information pour extraire les entités telles que les personnes, lieux, évènements ou la catégorisation de questions pour classer les questions en types connus (qui, où, quand, comment, etc.) » (GUPTA & LEHAL, 2009).

La huitième technique est le **suivi de thème** permettant de recommander des documents sur base de mots-clés déterminés par l'utilisateur ou par ses précédentes lectures. Le suivi de thème est utilisé dans différents domaines d'applications comme par exemple le domaine médical « quand les docteurs ou autres personnes regardent pour de nouveaux traitements de maladie et qu'ils souhaitent être au courant des dernières avancées » (GUPTA & LEHAL, 2009).

La dernière technique est **l'analyse des sentiments** (opinion mining)**.** Celle-ci vise à définir l'émotion, c'est-à-dire positive, négative, les deux ou neutre; uniquement sur base du vocabulaire utilisé dans les textes. Cette technique est particulièrement utilisée sur les réseaux sociaux tels que Twitter ou les avis sur un produit sur les sites internet pour analyser la satisfaction des clients.

Au vu de ce récapitulatif, les techniques les plus susceptibles de nous intéresser dans notre recherche sont :

- La visualisation de l'information qui serait utile afin de montrer les évolutions des techniques mais également les mots les plus utilisés.
- La modélisation de sujets afin de découvrir les domaines présents dans notre corpus et obtenir les mots les plus représentatifs de ceux-ci.
- Le regroupement pour regrouper les documents utilisant les mêmes mots et donc parlant d'un même domaine.

3.1.2 Pré-processing

Pour être appliquées de façon efficaces, toutes ces techniques de text mining doivent être précédées d'une phase de pré-processing car « les données textuelles contiennent souvent des

formats spéciaux comme les formats de nombres, dates, mots les plus courants qui sont peu susceptibles d'aider à l'exploration de textes tels que les prépositions, articles, pronoms qui peuvent être éliminés » (KANNAN, GURUSAMY, VIJAYARANI, ILAMATHI, & NITHYA, 2014). De plus, certaines données collectées ne sont parfois pas de bonne qualité pour la question de recherche, il faut donc retirer ces données afin d'éviter d'obtenir de mauvaises informations. La phase de pré-processing comprend 3 étapes : la collecte de données, le nettoyage des données et enfin la transformation. Tous les concepts de cette partie qui ont été utilisés dans la recherche seront illustrés dans la partie 4.3.

La collecte des données consiste à collecter différents textes qui seront analysés pour répondre à la question de recherche, plus communément appelé *corpora* ou *corpus de documents*. Les différents textes peuvent venir de multiples sources (web, documents d'entreprises, textes personnels, sondages, journaux, etc) et peuvent être de différentes tailles. Il est toutefois important que ces données soient en format digital afin de pouvoir appliquer les différentes techniques de text mining de manière automatisée.

Le nettoyage des données consiste à « supprimer les caractères non-importants, tokeniser le texte, convertir en minuscule, enlever les mots-vides et stemmatiser les mots » (KOBAYASHI, MOL, BERKERS, KISMIKÓH, & HARTOG, 2018). Les caractères non-importants sont multiples, il peut s'agir d'un espace, d'un hashtag, de nombres, etc. La tokenisation (segmentation) du texte permet de diviser le texte en phrases et/ou les phrases en mots. Les mots vides sont les conjonctions de coordination, les pronoms, les déterminants ou autres mots qui sont trop souvent utilisés dans les textes qu'ils n'apportent aucune information dans l'analyse. Il existe plusieurs listes de mots vides de référence en fonction de la langue. Les mots vides peuvent également être une liste de mots liés à un domaine d'application qui seront présents dans presque chaque document récolté et n'apportent donc pas d'information. Pour finir, la stemmatisation permet d'obtenir la racine d'un mot et donc de regrouper des mots sémantiquement similaires. Cela permet donc de réduire considérablement la taille du dictionnaire du corpus. Il existe aussi la lemmatisation. Ce concept est similaire à la stemmatisation car il permet d'obtenir un mot racine mais requiert de connaitre la partie du discours (« Part of Speech (POS)» en anglais). Cela indique si le mot est utilisé en tant que verbe, nom, adjectif, ou autres formes.

La troisième étape du pré-processing est la transformation. Celle-ci permet à l'ordinateur d'interpréter les données textuelles. En effet, celui-ci ne sait interpréter que des données

mathématiques, c'est pourquoi il est nécessaire de transformer le texte en chiffres. La transformation la plus utilisée est le modèle du vecteur. Chaque élément du vecteur représente le poids d'un terme (présent dans l'ensemble des documents) pour ce document précis. La taille du vecteur est donc la taille du vocabulaire présent dans l'ensemble des documents. Différentes représentations de vecteur existent, elles ont toutefois une efficacité différente en fonction de l'objectif à atteindre et ont chacune des avantages et inconvénients. Il y a tout d'abord le vecteur binaire ou One-Hot Encoding, où le poids est de 1 si le mot est présent dans le document ou 0 s'il ne l'est pas. Il y a également le vecteur de fréquence où le poids est le nombre de fois que le mot apparait dans ce document. Enfin, il y a la vectorisation TF-IDF (fréquence du terme * fréquence inverse du document), qui donne un poids plus important aux mots qui sont peu fréquents dans l'ensemble des corpus et fort présents dans le texte en lui-même. Un gros poids indique donc que ces mots sont importants dans l'interprétation du document. Cette méthode est la méthode la plus utilisée.

A ce pré-processing de base, certains rajoutent un pré-processing linguistique. Cela comprend la définition du POS de chaque mot, découpage du texte (« Text chunking »), la désambigüisation du sens d'un mot et l'analyse syntaxique. Le découpage du texte permet de grouper des mots qui sont souvent l'un à côté de l'autre dans une phrase. La désambigüisation permet d'obtenir la bonne signification d'un mot qui peut avoir plusieurs sens. Enfin, l'analyse syntaxique permet de faire des liens entre les mots. Toutefois, cela demande d'avoir accès à des lexiques qui contiennent des synonymes, antonymes ou autres (Par exemple Wordnet). Seulement, cela n'est pas encore assez développé et des recherches restent à faire. C'est pourquoi ce type de pré-processing n'est pas tout le temps utilisé et son utilisation dépend du contexte d'applications.

Une fois cette phase de pré-processing réalisée, les techniques du text mining peuvent être appliquées. Les résultats obtenus seront alors analysés avec la connaissance des utilisateurs pour en retirer l'information.

3.1.3 Applications des techniques de text mining à l'analyse de corpus scientifique

Il existe énormément d'articles scientifiques présentant les résultats de travaux menés dans différents domaines d'applications à l'aide de différentes techniques de text mining. Différents travaux sont réalisés sur différentes sources de données, que ce soit des articles scientifiques, des textes des réseaux sociaux ou encore des documents propres à un domaine d'application, etc. Pour les réseaux sociaux, il y a des travaux réalisés sur Twitter et Facebook

(SALLOUM S. A., AL-EMRAN, MONEM, & SHAALAN, 2017), Youtube (SEVERYN, MOSCHITTI, URYUPINA, PLANK, & FILIPPOVA, 2016). Pour les documents propres à un domaine d'application, il y a par exemple des décharges médicales (YANG, SPASIC, KEANE, & NENADIC, 2009), annonces d'emplois (PEJIC-BACH, BERTONCEL, MESKO, & KRSTIC, 2020), des sites d'évaluations en ligne ou blogs (KIM, PARK, YUN, & YUN, 2017).

Ces travaux sont réalisés dans un grand nombre de domaines d'applications différents, que ce soit la médecine, la finance, l'éducation, etc. Dans le cas de notre d'étude, nous allons nous intéresser plus particulièrement à 7 articles réalisés sur des articles scientifiques de divers domaines d'application. Nous axons également notre recherche sur des articles utilisant les méthodes susceptibles d'être utiles pour notre recherche, c'est-à-dire la modélisation de sujet, la visualisation et le regroupement (voir 3.1.1). Ceux-ci vont nous permettre de faire des comparaisons sur les méthodes de text mining utilisées pour l'analyse d'articles de recherche et les procédures à suivre. Le Tableau 3.1 représente une vue globale des différents articles analysés.

Comme nous avons vu dans le point 3.1.2, l'utilisation de techniques de text-mining nécessite une phase de pré-processing consistant à collecter les données, les nettoyer et les transformer. Nous allons analyser ces trois points dans la littérature.

L'analyse de la littérature permet de dire que la collecte des données se fait sur des bases de données spécialisées dans le domaine de recherche ou sur bases de données scientifiques générales. SUN & YIN (2017) justifient leur choix de journaux à l'aide d'index dans le domaine d'analyse, c'est-à-dire le transport. La collecte des articles se fait de manière automatisée sur base d'une liste de mots-clés identifiés par les auteurs ou des experts. Les articles collectés sont ceux qui contiennent ces mots-clés dans le titre, les mots-clés ou la description. La méthode de collecte varie d'une étude à l'autre et n'est pas toujours précisée.

Pour ce qui est du nettoyage, nous pouvons voir qu'il y a différents types de nettoyage. Tout d'abord, il y a un nettoyage des articles collectés. Notamment via l'analyse des mots qui ne sont pas pertinents pour leur recherche ((HAO, CHEN, LI, & YAN, 2018) et (MORO, PIRES, RITA, & CORTEZ, 2019)). Les derniers précisent que cette inspection se fait manuellement en analysant les titres et les descriptions. DELEN & CROSSLAN (2008) procèdent également à un nettoyage sur le type de revue en «supprimant les notes éditoriales, les notes de recherche et les aperçus exécutifs de la collection ». MORO, PIRES, RITA, &

CORTEZ (2019) ayant collecté les articles sur une base de données générales, ont ensuite gardé les articles qui ne provenaient que d'éditeurs connus afin d'assurer la pertinence des informations. Le nettoyage peut se faire aussi par la suppression d'articles sans contenu ou de courtes descriptions (moins de 10 mots) (SUN & YIN, 2017). Ensuite, la majorité des études procède à un nettoyage du texte via la mise en minuscule, la suppression de la ponctuation, l'enlèvement des mots-vides (stop-words), la stemmatisation. La liste des stop-words peut également être complétée par une liste propre au domaine (DELEN & CROSSLAN, 2008) ou d'un éditeur (ASMUSSEN & MOLLER, 2019). Ces derniers utilisent également un processus itératif pour valider le nettoyage manuellement car ils suggèrent que « le processus de nettoyage est terminé une fois que les articles chargés contiennent principalement des mots à valeur ajoutée » (ASMUSSEN & MOLLER, 2019).

Pour la transformation, la matrice terme-document est construite via des méthodes diverses. Certains utilisent la méthode binaire afin de pouvoir déterminer un seuil maximum ou minimum dans les documents. D'autres utilisent la fréquence afin de déterminer un seuil des mots qui reviennent trop rarement ou pas assez. Et enfin, d'autres utilisent la méthode TF-IDF afin de directement avoir les termes pertinents pour le document. DELEN & CROSSLAN (2008) et SALLOUM, AL-EMRAN, MONEM & SHAALAN (2018) réalisent la matrice sur base des descriptions de l'article alors que HAO, CHEN, LI, & YAN (2018) ; MORO, PIRES, RITA, & CORTEZ (2019) ; ASMUSSEN & MOLLER (2019) et SHARP, AK & HEDBERG JR (2018) utilisent l'intégralité des documents. Le choix de la description se justifie par le fait que « le résumé est une représentation compacte de l'article entier et il contient normalement suffisamment de mots clés sur les thèmes de recherche » ((STREYVERS & GRIFFITHS, 2004) cité dans SUN & YIN (2017)). DELEN & CROSSLAN (2008) disent qu'il n'est pas intéressant de s'attarder à la liste des mots-clés pour plusieurs raisons. Premièrement, car ils considéraient que ceux-ci se retrouveraient forcément dans la description et qu'il y a donc redondance d'information. Secondement, ils déterminaient que les mots-clés sont biaisés en pensant que les auteurs indiquaient ceux-ci en fonction des concepts auxquels ils voulaient que leurs articles soient reliés. Afin de réduire la taille de la matrice termes-documents qui peut être assez conséquente, un dictionnaire peut être utilisé (MORO, PIRES, RITA, & CORTEZ, 2019). Dans le cadre de leur problématique, ce dictionnaire est créé par des experts afin de déterminer les mots qui sont réellement liés au marketing et plus particulièrement le marketing ethnique. « La connaissance du contexte est un atout clé pour garantir le bon comportement de ces systèmes et doit être incluse pour

améliorer la précision des systèmes » (MORO, PIRES, RITA, & CORTEZ, 2019). HAO, CHEN, LI, & YAN (2018) et ASMUSSEN & MOLLER (2019) quant à eux utilisent un seuil de fréquence en pourcent qu'ils ont déterminé sur base d'une analyse manuelle pour réduire la taille de la matrice en enlevant les termes trop fréquents ou les termes trop rares sur base d'une analyse des fréquences présentes dans leur matrice. SUN & YIN (2017) font de même mais avec une valeur absolue plutôt qu'un pourcentage. Afin d'obtenir une matrice de dimension inférieure, SHARP, AK & HEDBERG JR (2018) utilise la méthode LSA qui permet de regrouper des termes liés en concept et de lier ces concepts aux documents. D'autres encore utilisent la décomposition en valeur singulière afin de réduire la matrice (DELEN & CROSSLAN, 2008).

Du point de vue des outils utilisés pour réaliser les analyses de textes, la majorité utilise le langage de programmation R avec ses différents modules qu'il propose. D'autres utilisent des logiciels fournis par des entreprises afin de faire de la science des données et le stockage des données tels que RapidMiner ou Microsoft access. D'autres encore utilisent le langage de programmation Python dont le module nltk est utile pour le traitement texte.

Si nous nous intéressons maintenant aux techniques de text mining utilisées, nous pouvons voir que les méthodes utilisées sont diverses.

Tout d'abord, nous pouvons observer que tous ceux qui utilisent la modélisation de sujets le font via la méthode LDA. Cependant, il n'y a pas de méthode précise pour obtenir le nombre de sujets qu'il faut spécifier dans le modèle LDA. MORO, PIRES, RITA, & CORTEZ (2019) ont tout d'abord fixé un nombre et puis ont diminué petit à petit ce nombre jusqu'à ce qu'ils trouvent un nombre qui semblait correct avec un bon niveau de groupement des documents. HAO, CHEN, LI, & YAN (2018) et ASMUSSEN & MOLLER (2019) ont évalué le nombre de sujet optimal sur base de la perplexité obtenu pour chaque nombre. « Un score faible indique un meilleure modèle de généralisation » (ASMUSSEN & MOLLER, 2019). Ils précisent toutefois que le nombre optimal de sujets dépend également du contexte. L'idéal est de « trouver l'équilibre entre un nombre utilisable de sujets et, en même temps, de garder la perplexité aussi faible que possible » (ASMUSSEN & MOLLER, 2019). Afin d'assurer une validation dans les données, ils ont tous les deux utilisé une validation croisée. ASMUSSEN & MOLLER (2019) ont également testé le modèle LDA en modifiant d'autres paramètres que le nombre de sujets optimal, c'est-à-dire le temps de rodage, le nombre d'itération, la valeur de départ, le nombre de plis, et la distribution entre les ensembles

d'entrainements et de tests. Une fois le nombre de sujets optimal trouvé et les paramètres définis, le modèle LDA est appliqué à l'ensemble des données. Les résultats obtenus avec le modèle LDA peuvent être plus ou moins précis selon le contexte d'étude. SUN & YIN (2017) précisent que certains sujets obtenus peuvent être liés à des domaines de recherches alors que d'autres sont trop généraux et ne peuvent donc pas être associés à un domaine précis. ASMUSSEN & MOLLER (2019) indiquent également qu'il n'est pas toujours possible de donner une étiquette à chaque sujet découvert et que cela dépend du choix optimal du nombre de sujets. Des études doivent encore être réalisées à ce point de vue là.

Ensuite, pour ce qui est du regroupement, différentes méthodes ont été utilisées selon le cadre d'étude. SALLOUM, AL-EMRAN, MONEM & SHAALAN, K. (2018) utilisent la méthode traditionnelle k-means. Le regroupement est utilisé par HAO, CHEN, LI, & YAN (2018) dans le but de regrouper les sujets de modèles similaires obtenus sur bases de l'analyse des sujets. Ils utilisent un regroupement hiérarchique une fois avec la mesure de similarité entre les termes des sujets et une autre fois avec similarité entre les documents des sujets. Dans les deux cas, la similarité est basée sur la similarité du cosinus. DELEN & CROSSLAN (2008) utilisent le regroupement avec un algorithme de maximisation des attentes. SHARP, AK & HEDBERG JR (2018) ont effectué un clustering flou afin de regrouper les documents similaires. Pour faire cela, ils ont utilisé la similarité du cosinus entre les documents.

Pour la visualisation, plusieurs objets sont utilisés dans la littérature : les nuages de mots, les diagrammes, des graphes. Certains utilisent même des cartes géographiques lorsqu'ils présentent des informations par région (HAO, CHEN, LI, & YAN, 2018). Une contribution du travail réalisé par MORO, PIRES, RITA, & CORTEZ (2019) est « la présentation des sujets dans une nouvelle image visuellement attrayante (carte des sujets) qui augmente considérablement la lisibilité et l'interprétation par rapport aux tableaux complexes utilisés dans la littérature existante».

Pour ce qui est de l'analyse des résultats, la plupart effectue une analyse des mots les plus fréquents. De plus, il y a toujours une analyse des termes et des documents présents dans chaque groupe obtenu, que ce soit via le regroupement ou la modélisation de sujets. HAO, CHEN, LI, & YAN (2018) fournissent également des analyses descriptives des données avec les publications par année, les sources de publications productives, la distribution géographique ou encore les auteurs et institutions productifs. SUN & YIN (2017) mettent surtout l'accent sur les analyses dans le temps et par région.

En conclusion, à travers la revue de la littérature, nous avons pu observer que la technique de modélisation de sujets était plus pertinente dans le cadre d'articles scientifiques afin d'obtenir les sujets présents dans le corpus. La méthode la plus utilisée dans la littérature est LDA. Toutefois, il faut garder à l'esprit que tous les articles présentés ici se focalisent sur un domaine particulier et donc que les sujets recherchés sont des sous-domaines et ont un lien hiérarchique. Or dans notre étude, les sujets que nous nous attendons à trouver n'ont pas de domaine hiérarchique commun. La méthode utilisée peut donc différée ou nous pouvons obtenir des résultats de moindre qualité. Il a également été observé que la méthode LDA est une méthode subjective et que la modélisation de texte requiert une connaissance du domaine afin d'obtenir de meilleurs résultats et donc une intervention humaine. Une phase importante pour obtenir de bons résultats est également le nettoyage. Ce nettoyage s'effectue en deux phases : analyse des articles/informations pertinents et pré-processing sur le texte. Nous nous attendons donc à passer une grande partie de notre analyse sur ce nettoyage afin de maximiser la probabilité d'avoir de bons résultats. Afin de maximiser l'interprétation des résultats, l'utilisation d'outils visuels tels que des graphiques, nuages de mots, etc est importante.

	Base de données	Attributs	Nettoyage	Transformation	Outil	Méthode utilisée	Analyses
(SALLOUM S. , AL-EMRAN, MONEM, & SHAALAN, 2018) : Apprentissage mobile	Springer, Wiley, Science Direct, SAGE, IEEE, and Cambridge	Mots-clés Années de publication Source intégralité du contenu	Bruit linguistique Stop-words Minuscule, Longueur du token (4-25)	Fréquence	RapidMiner	Nuage de mots, Diagrammes Règles d'association Mesure de similarité Clustering avec kmeans	Fréquence des mots-clés à travers toutes les BDD et par BDD indivuellement Coocurrence des mots Similarité entre les documents Groupes de sujets Evolution des articles par années
(HAO, CHEN, LI, & YAN, 2018) : Médecine	Web of Science, PubMed	Titre Auteurs Année Source Description Adresse des auteurs Pays de l'institution	Supprimer articles avec mots non pertinents	TF-IDF, seuil de 0.1 avec analyse manuelle	R	Analyse réseaux sociaux auteurs LDA perplexité pour k optimal, Clustering hiérachique sur groupes avec similitude cosinus Diagramme, carte, graphe	Publications par année Source de publications productives Distribution géographique Auteurs et institutions productifs Collaboration Termes les plus fréquents Groupes de sujet Tendances sujets
(MORO, PIRES, RITA, & CORTEZ, 2019) : Marketing ethnique	Scopus	Titre Mots-clés Description Intégralité du contenu	Editeurs connus Articles non pertinents Stemmatisation	fréquence + utilisation d'un dictionnaire pour réduire la taille de la matrice	R	LDA → multiple test pour trouver k optimal Nuage de mots	Sujets et termes présents dans ceux-ci
(DELEN & CROSSLAN, 2008) : Système de gestion d'informations	MIS Quarterly (MISQ), Information Systems Research (ISR) et le Journal des systèmes d'information de gestion (JMIS).	Titre Description Auteurs Mots-clés Volume Numéro Année	Stops-words spécifiques au domaine + stops-words courants Stemmatisation Synonymes Types de revue	Fréquence + Décomposition en valeur singulière pour réduire matrice	Microsoft Access	Clustering avec un algorithme de maximisation des attendes	Fréquence mots par période de 2 ans pour toutes les BDD et individuellement Termes les + présents dans groupe Groupes : répartition par BDD et par année

(SHARP, AK, & HEDBERG JR, 2018) : Production d'entreprise	Primarily Engineering Village and Google Scholar	**Description**	Mots-vides, Ponctuations	TF-IDF + approximation avec LSA		Clustering flou avec similarité cosines Visualisation avec diagrammes	Sujets centraux de chaque groupe Termes les plus fréquents
(ASMUSSEN & MOLLER, 2019) : Approche pratique de la modélisation de sujets pour la revue exploratoire de la littérature		**Intégralité du contenu**	Minuscule Ponctuation Informations non pertinentes Stops-words d'un éditeur + courants Processus itératif	Fréquence Binaire + réduction avec mots dans 99% des doc ou rare	R	LDA (perplexité avec k folds (75% -25%) + petit nombre de sujet car on veut globalité Quand k trouvé: appliquer LDA à l'ensemble Diagrammes, tableaux	Mots les plus courants par sujets Sous-thèmes
(SUN & YIN, 2017) : thèmes et tendances dans la recherche sur les transports	Journaux choisis sur base de Science Citation Index et Social Science Citation Index	**Description** Auteurs Années Pays Source	Sans contenu Description trop courte Token Stop-word de nltk	Fréquence Binaire pour réduction avec moins de 5 documents Fréquence pour réduction de plus de 6000 apparitions	Python NLTK	LDA, k choisi sur base d'analyse subjective des résultats Nuage de mots, diagrammes, graphes	Distribution articles par journaux et évolution dans le temps Thèmes par sujet Evolution dans le temps des sujets pour tous les journaux et par journal Distribution des sujets de revues Similitude des revues Distribution des thèmes par région Co présence de mots

Tableau 3.1: Résumé revue de la littérature

3.2 Choix du journal

Maintenant que nous avons un aperçu des techniques de text mining et de la phase de pré-processing, nous pouvons commencer à réaliser notre étude. Pour rappel, celle-ci vise à déterminer l'évolution des usages de techniques de machine learning dans différents domaines d'applications. Cette analyse se fait via des techniques de text mining sur différents articles scientifiques. Avant de commencer notre analyse, il est important de choisir le journal dans lequel les articles scientifiques seront collectés. Ce choix s'est porté sur le journal ***Expert Systems with Applications***. Celui-ci a été choisi en tenant compte de plusieurs critères.

Tout d'abord, pour obtenir des informations pertinentes, il fallait choisir un journal de qualité selon des critères scientifiques. Pour ce faire, différents classements réalisés par *Scimago and Country rank*[5] ont été analysés. La première analyse s'est faite par rapport au classement sur base de l'indice h (Annexe 1). L'indice h permet de déterminer l'impact scientifique d'un journal. L'indice h « tient compte de la productivité (nombre d'articles publiés) et de l'impact (nombre de citations reçues) en comptant les citations les plus employées d'un chercheur ainsi que le nombre de citations que ces œuvres ont reçues dans d'autres publications » (CENTRE UNIVERSITAIRE DE SANTE DE Mc GILL, 2021). Selon ce classement, il est possible de voir que ce journal se situe en cinquième position en 2019 avec une valeur de 184. Cela permet d'indiquer que c'est un journal de qualité et donc que les articles de recherches qu'il contient sont fortement susceptibles de nous donner de l'information pertinente. Un autre point important est que le journal *Expert Systems with Applications* est également bien situé dans le classement réalisé sur base du critère de Scimago Institutions Rankings. L'organisme effectue ce classement en prenant en compte que chaque citation n'a pas la même valeur. Via ce critère, le journal se situe en 33ème position sur 602 journaux avec une valeur de 1.494. Les bonnes positions de ce journal dans ces deux classements permettent donc de dire que c'est un journal fiable et de qualité.

Un autre critère pour le choix du journal est qu'il fallait un journal qui s'appliquait à plusieurs domaines au vu de la question de recherche. Au vu du titre et de la thématique du journal *Expert Systems With Application*, ce journal s'applique à plusieurs domaines. C'est en effet le cas puisque la description du journal indique qu'il publie des articles dans différents domaines comme la finance, l'ingénierie, la médecine et encore bien d'autres.

[5] Organisme permettant de trier les journaux selon différents critères.

Le dernier critère du choix du journal est lié à la période d'analyse de la question de recherche. En effet, puisqu'il s'agit d'étudier l'évolution des techniques de machine learning il était donc nécessaire d'avoir un journal qui couvre une période assez longue. *Expert Systems with Applications* permet d'obtenir les publications pour 30 années de 1990 à 2020.

Malgré que quatre journaux aient un meilleur indice h que le journal *Expert Systems with Applications*, ils n'ont pas été choisis pour plusieurs raisons. Les trois premiers (*IEEE Transactions on Pattern Analysis and Machine intelligence, IEEE on Neural Networks and Learning Systems* and *Pattern recognition*) sont des journaux qui ne s'appliquent qu'à certaines techniques de ML comme les réseaux de neurones ou la reconnaissance de modèles. Ils n'étaient donc pas pertinents pour répondre à la question de recherche qui est liée à l'ensemble des différentes techniques de ML. Le quatrième journal *Journal of machine learning research* est quant à lui trop orienté vers l'aspect théorique des techniques de machine learning et non orienté vers la mise en pratique de ces techniques dans différents domaines d'applications. De plus, il couvre une période plus limitée (2001-2021). C'est donc pour ces différentes raisons, que ces quatre journaux n'ont pas été pris en compte.

En conclusion, étant donné que le journal *Expert Systems with Applications* correspond à tous les critères de recherche, il semble pertinent d'utiliser ce journal. De plus, il fournit un bonus dans le cadre de notre recherche étant donné qu'il est accessible aux étudiants de l'Université de Namur via la Bibliothèque Universitaire Moretus Plantin. Cet accès est vu comme un bonus puisqu'il permettra d'aller lire certains articles en profondeur si cela est nécessaire pour comprendre les résultats obtenus.

3.3 Processus

Afin d'analyser les évolutions des techniques de machine learning dans différents domaines d'applications, il est nécessaire de suivre un processus stricte (Figure 3-1). Ce processus a pu être établi suite à l'analyse des études menée sur l'application de text mining sur les articles scientifiques (partie 3.1.3).

Les quatre premières étapes concernent la collecte et le nettoyage des données afin d'obtenir des articles dont le sujet concerne le machine learning, proviennent du journal *Expert Systems with Applications*. Ces étapes permettent aussi d'obtenir des attributs de qualité et éviter des biais. Toutes ces étapes seront expliquées en détail dans la partie 0, préparation des données. Comme cela a déjà été fait dans certaines études de text mining sur des articles scientifiques, il a été décidé de faire une analyse descriptive afin d'obtenir une première information

concernant nos différents articles collectés. Cette analyse sera expliquée dans la partie 5 « Analyses et résultats ». Ensuite, un tri sera fait sur les articles afin d'écarter les articles uniquement théoriques et ne garder que ceux qui sont applicatifs. Cette phase sera également expliquée dans la partie 4 « préparation des données ». La phase suivante est le pré-processing sur les descriptions des articles afin de pouvoir ensuite appliquer une technique de text mining dessus. Cette étape sera également expliquée dans la partie 4 « préparation des données ». Une fois la préparation des données faites, nous pouvons appliquer les techniques de text mining. Au vu des différentes techniques analysées dans la littérature, il a été décidé d'appliquer les techniques de la modélisation des sujets et de la visualisation. La première permettra de faire ressortir des domaines présents dans notre base de données. La deuxième permettra quant à elle une meilleure interprétation des résultats, plus particulièrement pour l'évolution dans le temps qui est plus compréhensive avec un graphique évolutif. Pour finir, nous pourrons analyser l'évolution des techniques de machine learning au sein de ces différents domaines trouvés. Les résultats seront présentés dans la partie 5.

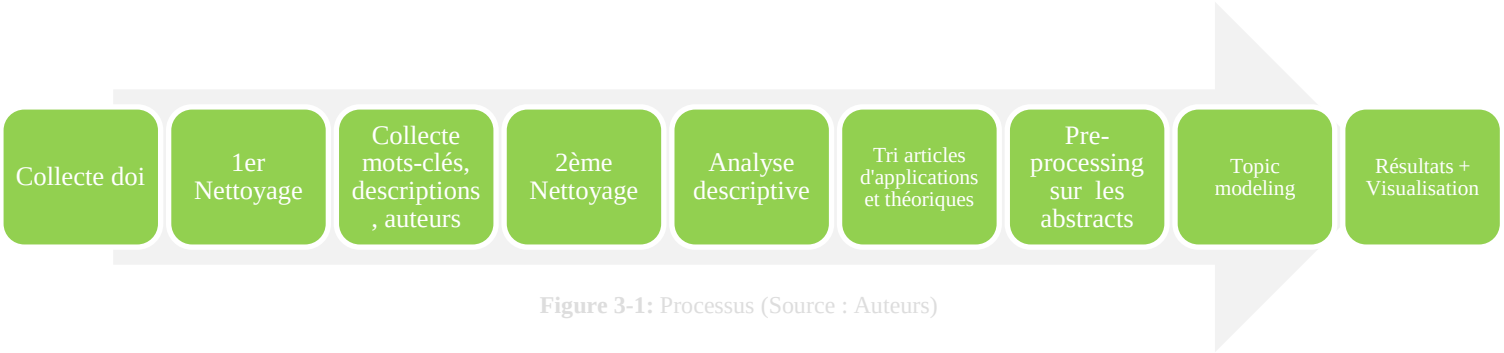


Figure 3-1: Processus (Source : Auteurs)

4 Préparation des données

4.1 Collecte et nettoyage

4.1.1 Collecte des articles

Les données collectées sont des articles scientifiques du journal *Expert Systems with Applications* publiés par l'éditeur ELSEVIER (2021) sur le site ScienceDirect. Ce site publie des articles sur différents domaines et est à la pointe de l'évolution. Elsevier fournit une API (application programming interface) afin de permettre aux chercheurs, gouvernements, professeurs, élèves et autres personnes intéressées par la recherche, de pouvoir collecter des données rapidement et en masse.

Pour pouvoir faire cette collecte de données en masse, il a fallu obtenir une clé API en se créant un compte Elsevier. Cette clé a ensuite été utilisée dans le code de programmation afin de collecter tous les documents souhaités de manière automatisée sur base de mots-clés. Dans cette étape, il y a donc deux points. Tout d'abord définir les mots-clés qui seront utilisés pour la collecte. Ensuite déterminer la requête afin de collecter automatiquement les articles.

4.1.1.1 Choix des mots-clés

La création du corpus de documents s'est faite sur base des mots-clés suivant : ***machine learning, data mining, learning algorithm, decision tree, support vector machine, artificial intelligence, learning method, random forest, neural networks, supervised learning*** et ***unsupervised learning***. Ce choix de mots-clés a été pris sur base des termes reliés au premier niveau ou au deuxième niveau au terme *machine learning* selon Science Direct tout en gardant également un regard critique vis-à-vis de la question de recherche. Science Direct détermine « les termes reliés à un mot sur base de techniques de ML utilisées sur un ensemble de documents et permettant d'extraire de l'information » (SCIENCE DIRECT, 2021). Ces techniques permettent d'obtenir une définition du terme principal, les mots reliés et un ensemble d'articles qui sont pertinents pour ce mot. Les termes reliés de premier niveau au mot « machine learning » sont donc les mots directement liés au terme *machine learning* comme par exemple *data mining* ou *neural networks* (Annexe 2). Les termes de deuxième niveau sont des termes indirectement liés au terme machine learning via les termes de premier niveau. Cela permet d'avoir un niveau plus détaillé concernant le concept du ML. Un exemple est le mot *decision tree* qui est lié à *learning method*, qui était un contact de premier niveau. Un regard critique vis-à-vis de ces mots de premier niveau et deuxième niveau est nécessaire

afin d'éviter de collecter des articles via plusieurs mots-clés. Par exemple, le mot *artificial neural network* n'a pas été utilisé, alors qu'il était un mot de premier niveau, car il est étroitement lié au mot *neural networks* et donc il y aurait eu beaucoup de redondances dans les données collectées.

4.1.1.2 Requête

La collecte des données a été réalisée de manière automatisée via une implémentation dans le langage de programmation Python. Ce choix de langage s'est fait étant donné qu'il existe un module python *elsapy* utile pour la collecte d'articles scientifiques. Ce module a pour but « de faciliter la vie des personnes qui ne sont pas principalement des programmeurs, mais qui ont besoin d'interagir avec les données de publication et de citation des produits Elsevier de manière programmatique (par exemple les chercheurs universitaires)» (GITHUB, 2021). La collecte des informations s'est réalisée en trois étapes.

La première étape avait pour but de se connecter avec la plateforme Elsevier. Pour ce faire, nous avons utilisé la fonction *ElsClient* du module *elsapy* avec comme paramètre la clé API qui a été fournie par Elsevier lors de la création d'un compte. C'est grâce à cette clé API que la connexion à la plateforme Elsevier se fait.

La deuxième étape était de collecter les articles et certaines de leurs informations qui nous permettront de répondre à notre question de recherche. Pour ce faire, nous avons utilisé la fonction *ElsSearch* du module *elsapy*. Cette fonction nécessite une requête qui va être appliquée à la base de données Elsevier afin de fournir les informations. La structure de la requête utilisée se trouve à la Figure 4-1. Celle-ci indique que tous les documents qui ne sont pas des conférences ou comptes-rendus, qui ont comme titre de journal *Expert Systems with Application* et qui contiennent le mot-clé *keywords* dans le titre, l'abstract ou les mots-clés pour l'année *year* seront collectés. Cette requête sera répétée pour chaque paire de mots-clés et années intéressante dans le cas de notre recherche, à savoir les mots-clés cités dans la partie 4.1.1.1 et les années de 1990 à 2020. Chaque donnée collectée contient plusieurs attributs. Ceux qui nous intéressent plus particulièrement sont les titres, la date, le type de publication, l'identifiant.

```
query = "TITLE-ABS-KEY({0}) AND SRCTITLE((Expert+Systems+with+Applications) AND NOT (Proceedings) AND NOT(IEEE) AND NOT(Conference)) AND PUBYEAR = {1}".format(keywords,year)
```

Figure 4-1: Structure de la requête (Source : Auteurs)

4.1.2 Premier nettoyage des articles

Pour des questions d'efficacité, d'optimisation de stockage et de pertinence des informations, plusieurs nettoyages de données sont réalisés à différentes étapes du processus. Le premier nettoyage a eu lieu après la première collecte de données c'est-à-dire la collecte des articles et de leurs identifiants sur base de différentes requêtes. Ce nettoyage porte sur différents problèmes détectés dans les données. Ceux-ci sont les doublons, les données non pertinentes provenant d'un autre journal (*Expert Systems with Applications: X*), des données autres que des articles, les attributs non pertinents et le manque d'identifiant.

Pour éviter le risque de duplication d'informations, liée au fait que plusieurs articles ont été collectés via différents mots-clés et stockés à différents endroits, une concaténation des différentes bases de données a été faite. Les doublons ont été effacés pour ne plus avoir d'informations redondantes.

Le problème de données non pertinentes provenant d'un journal autre que celui utilisé dans notre recherche vient d'un manque de précision dans la requête. Pour remédier à cela, il a fallu procéder à un filtrage de la colonne *publicationName* pour n'obtenir que les éléments qui indiquent *Expert Systems with Applications.* Afin de remédier aux problèmes des données autres que des articles, également liés à un manque de précision de la requête, un filtrage a été fait sur la valeur *Article* pour l'attribut *subTypeDescription*.

Les attributs non pertinents étaient multiples ; la numérotation de la page de l'article dans le volume, l'affiliation, etc. Certaines informations se retrouvaient même dans deux attributs, par exemple le lien internet qui se trouvait à la fois dans l'attribut *link* et *url*. Toutes ces colonnes ont été supprimées de la base de données.

Les articles sans identifiants étaient un gros problème car pour collecter les mots-clés, la description et les auteurs d'un article, il était absolument nécessaire d'avoir cet identifiant *doi*. Etant donné que le nombre de lignes sans identifiant ne représentait qu'une faible part de toutes les lignes collectées, il a été décidé de les supprimer car la perte d'information n'était pas conséquente.

A la fin de ce nettoyage, nous avons 5828 articles pour lesquels nous pouvions collecter les mots-clés, la description et les auteurs.

4.1.3 Collecte des mots-clés, descriptions et auteurs

Une fois le premier nettoyage réalisé, la troisième étape de la collecte des données a pu être réalisée. Cette collecte était nécessaire étant donné que l'attribut *auteur* ne fournissait qu'un auteur. Cela semblait étrange puisque les articles scientifiques sont souvent une coopération entre plusieurs chercheurs. Cette deuxième collecte de données a été réalisée pour collecter les auteurs mais également les mots-clés et les descriptions des articles, qui seront utiles pour l'analyse des données.

Afin de réaliser cela, la fonction *FullDoc* du module *elsapy* a été utilisée. Les auteurs, descriptions et mots-clés ont été récoltés sur base de l'identifiant de chaque article. Comme attendu, la fonction *FullDoc* nous a fourni une liste complète des auteurs de l'article.

4.1.4 Second nettoyage

Après avoir collecté les descriptions, auteurs et mots-clés, un nouveau nettoyage des données adresse plusieurs problèmes:

- Les mots clés et auteurs sont stockés sous forme de chaines de caractères, ce qui n'est pas pratique pour faire des analyses dessus.
- Il existe plusieurs formulations pour un même mot-clé ou des synonymes, dû au libre choix des auteurs de choisir les mots qu'ils souhaitent.
- Informations manquantes pour certains articles.

Nous allons voir plus en détails ces problèmes ainsi que le traitement nécessaire et certains résultats obtenus.

4.1.4.1 Mots-clés / auteurs sous forme de chaine de caractères

L'ensemble des mots-clés liés à un article sont stockés sous forme de chaines de caractères (String). Cependant, il n'est pas évident d'itérer sur ce type de données. Cela est problématique puisque l'itération sur les mots-clés sera souvent sollicitée dans le cadre de notre recherche.

C'est pourquoi il était nécessaire de transformer cette chaine de caractères en une liste de mots-clés, permettant une meilleure manipulation des données.

Ce nettoyage commence par l'enlèvement de certains caractères spéciaux comme « []' " » ainsi que les sauts à la ligne, les doubles espaces, les espaces en début et fin de mot et la transformation en liste.

Le même nettoyage a été appliqué à l'ensemble des auteurs.

4.1.4.2 Mots-clés : synonymes et formulations multiples

Etant donné que l'auteur est libre dans le choix de ses mots-clés, il existe de multiples formulations (abréviations/mots en entiers, pluriels/singulier, détaillés/non détaillés, minuscule/majuscule) pour un même mot ainsi qu'un ensemble de synonymes. Un exemple serait que certains auteurs notent *support vector machine* alors que d'autres notent *svm*. Un autre exemple est que certains précisent le mot *aco algorithm*, alors que d'autres utilisent le terme plus large *aco*.

Cela nécessite d'analyser l'ensemble des mots-clés et de trouver les mots de même signification mais étant écrits de manières différentes.

Afin de réduire ces différences dans la manière d'écrire des auteurs, plusieurs choses ont été mises en place. Tout d'abord, les mots ont été mis sous forme minuscule. Sur base de cela, nous avons obtenu une liste d'environ 13500 mots-clés. Afin de réduire la longueur de cette liste, nous avons procédé à une analyse manuelle pour déterminer les termes semblables. L'utilisation d'une fonction de mesure entre deux chaines de caractères a permis de faciliter cette analyse manuelle. Cette analyse nous a permis de déterminer des mots de même signification mais de styles différents. Par exemple, *support vector machine* et *svm*. Cette méthode reste toutefois subjective étant donné qu'elle n'a pas été réalisée de manière automatisée sur base d'un dictionnaire scientifique. Il existe le thésaurus Wordnet qui fournit des synonymes, antonymes et qui est utilisé dans plusieurs modules *python* pour l'analyse de texte. Toutefois, ce module ne va pas à un niveau assez détaillé dans les domaines d'applications. Il ne contient que les mots souvent utilisés dans la vie courante. Il n'était donc pas intéressant de l'utiliser puisque le corpus contenait trop de mots spécifiques au domaine machine learning. Les synonymes ont été choisis sur base de recherche sur internet et de connaissance. Cela a toutefois permis de réduire le niveau de subjectivité liée à la façon d'écrire des auteurs étant donné que les mots-clés sont maintenant déterminés sur base d'une seule personne.

Le Tableau 4.1 montre quelques exemples de mots-clés avant et après nettoyage. Cela permet de voir que de 6 mots-clés, notre nettoyage a permis de réduire à 2 mots clés.

Mot-clé de base	Mot-clé utilisé
least square support vector machine	lssvm
least square support vector machine lssvm	lssvm
least squares support vector machine	lssvm
artificial neural network ann	artificial neural network
artificial neural networks anns	artificial neural network
artificial neural networks ann	artificial neural network

Tableau 4.1: Modification mots-clés (Source : Auteurs)

4.1.4.3 Informations manquantes

Pour finir, plusieurs articles avaient des données vides pour les mots-clés ou pour les descriptions. Au total 210 articles étaient sans mots-clés, alors qu'un seul article n'avait pas de description. Comme il est possible de voir sur la Figure 4-2, les articles sans mots-clés se situent surtout avant les années 2000. Cela est certainement dû à un changement de politique concernant la publication des articles scientifiques et à l'obligation de mettre des mots-clés.

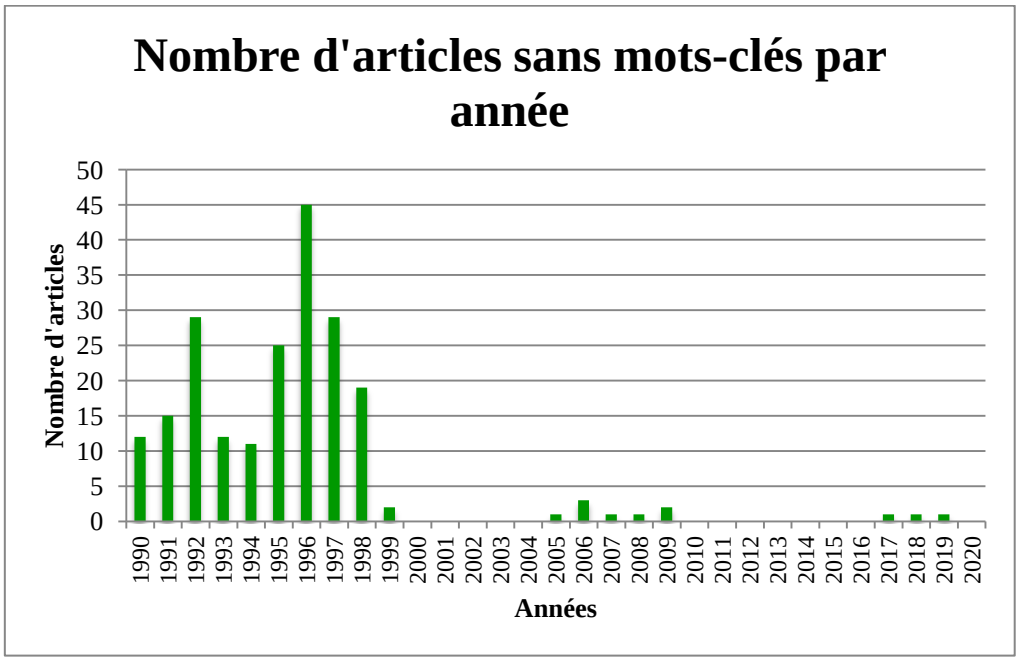


Figure 4-2: Nombre d'articles sans mots-clés (Source : Auteurs)

Au vu des ces informations manquantes, deux choix étaient possibles : les enlever ou générer ces informations manquantes. Etant donné qu'il n'y avait qu'un article sans description, il a été décidé de le supprimer. Par contre, le nombre d'articles sans mots-clés étant conséquent, la suppression de ces articles impliquerait une grosse perte d'information concernant les années 90. C'est pourquoi, il a été décidé de générer des mots-clés pour ces articles plutôt que de les supprimer.

Pour générer les mots clés, une tokenisation a été réalisée sur la description de ces articles. Nous avons ensuite calculé la fréquence inversée TF-IDF de chaque chaine de caractères composée de deux mots ou d'un seul. Nous avons ensuite décidé de prendre comme potentiels mots-clés les mots qui se situaient déjà dans la liste totale des mots-clés car nous savons que s'ils ont été utilisés, ils ont de forte chance d'être pertinent. Pour finir, nous avons choisi comme mots-clés, les cinq meilleurs en termes de fréquence TF-IDF. Tout d'abord, les meilleurs pour les chaines de caractères comprenant deux mots puis pour les mots uniques pour finalement obtenir une liste de cinq mots. Ce procédé a été choisi puisque la plupart des mots-clés sont souvent des mots doubles et non uniques. Un exemple de mots-clés générés par ce processus se trouve à l'annexe 3. Grâce à ce procédé, il n'y a plus un seul article sans mots-clés.

L'analyse descriptive sera faite après ce point de nettoyage et comprendra donc les articles liés à la théorie du ML et les articles d'applications.

4.2 Tri des articles théoriques et applicatifs

Etant donné que le journal *Expert Systems with Applications* publie également des articles qui ne sont axés que sur la théorie du ML, il était nécessaire de supprimer ces articles avant d'appliquer les techniques de text mining pour n'avoir dans les sujets que des domaines d'applications.

Afin de trier les articles, nous avons procédé à un tri sur les différents mots-clés afin de déterminer s'ils étaient susceptibles de représenter un domaine ou non. Pour se faire, nous avons commencé par retirer tous les mots ou abréviations représentant des techniques de ML et mots liés à l'intelligence artificielle publiés sur les pages "Outline of machine learning, Glossary of artificial intelligence et Glossary of computer science " de Wikipédia. Cela a permis de réduire la liste de mots-clés à trier. L'identification de mots-clés représentant un domaine s'est ensuite faite manuellement sur base de connaissances et recherches sur internet. Des mots tels que « disease, stock, credit » étaient considérés comme appartenant à un domaine. Une fois ce tri réalisé, nous avons pu analyser si les documents étaient uniquement liés à la théorie ou non. Si les documents avaient au moins 1 mot-clé représentant un domaine, alors ils étaient considérés comme applicatifs. Dans l'autre cas, ils étaient considérés comme théoriques.

Ce nettoyage a permis d'obtenir 3687 articles applicatifs contre 2140 articles théoriques.

4.3 Pré-processing

Comme vu dans la revue de la littérature, il faut une phase de pré-processing, qui consiste à collecter les données, les nettoyer et les transformer, avant d'appliquer les techniques de text mining. Dans le cadre de la recherche, les données sur lesquelles nous allons appliquer une technique de text mining sont les descriptions (abstract). La collecte de ces descriptions a déjà été expliquée dans la partie 3.1, c'est pourquoi ce point n'abordera que le point de nettoyage et de transformation des descriptions. Tous les points ont été réalisés à l'aide de différents modules provenant du package NLTK[6] en *python*.

4.3.1 Tokenisation

La tokenisation s'est réalisée à l'aide du module *tokenize*. Cela a permis d'obtenir la description (chaine de caractères) en une liste de mots simples appelés token. Cette séparation s'est faite sur base des espaces et de certaines ponctuations. Chaque token a ensuite été nettoyé, c'est-à-dire que les caractères spéciaux, chiffres et espaces ont été enlevés et que tout a été mis en minuscule.

4.3.2 Stop-words

NLTK fournit également une liste de mots-vides (*stop-words*). Les mots-vides sont des mots couramment utilisés dans des phrases et qui n'apportent pas d'informations. Ce sont donc les conjonctions de coordination, des verbes à forte utilisation, des déterminants, des adverbes, etc. Cette liste de mots-vides a été complétée par une liste de mot-vides fournit par Elsevier, pour encore plus de pertinence. La liste totale des mots-vides se trouve à l'annexe 4. Chaque token présent dans la liste des mots-vides a été retiré de la liste des tokens représentant la description. A la fin de ce processus, il n'y avait plus qu'une liste de tokens susceptible d'apporter de l'information pour la recherche.

4.3.3 Lemmatisation et stemmatisation

Lors de la phase de pré-processing, un choix s'impose entre la stemmatisation et lemmatisation. La première vise à obtenir la racine du mot c'est-à-dire que les préfixes, suffixes et pluriels des mots vont être enlevés sur base de certaines règles. Différents algorithmes existent pour réaliser cette stemmatisation. Le module NLTK en propose 2 ; Porter et Snowball. Le deuxième est une amélioration du premier. Pour ce qui est de la lemmatisation, cela consiste à prendre la racine du mot tout en tenant compte de la position

[6] NLTK est une plateforme de premier plan pour la création de programmes Python pour travailler avec des données en langage humain. (NLTK PROJECT, 2021)

dans la phrase. La lemmatisation de NLTK se base principalement sur Wordnet[7] et va chercher les liens entre les mots grâce à cette base de données. Un exemple de stemmatisation est que le mot « better » avec la position adjectif sera transformé en « good ». Un choix a donc dû être fait entre ces différentes propositions pour obtenir la racine d'un mot. Ce choix s'est porté sur la stemmatisation avec Snowball car la stemmatisation se fait avec Wordnet. Cependant, Wordnet ne va pas à un niveau assez précis des mots spécifiques à un domaine, mais est plutôt un lexique de mots courant de la vie. Cela n'était donc pas pertinent dans notre cas car beaucoup de mots sont vraiment spécifiques à un domaine, que ce soit au machine learning ou des termes médicaux, financiers, etc. Il y aurait donc eu énormément de mots qui n'auraient pas été mis à une forme racine. Snowball a été préféré à Porter étant donné que c'est une amélioration de celui-ci.

Le tableau ci-dessous (Tableau 4.2) montre un exemple de description qui est passée à travers le processus de nettoyage. Il est possible de voir que la stemmatisation a permis de réunir le mot ***traditional*** et ***tradition*** à la même racine ***tradit***, ce qui est donc efficient car il signifie la même chose.

Description	Token
Machine fault diagnosis is a traditional maintenance problem. In the past, the maintenance using tradition sensors is money-cost, which limits wide application in industry. To develop a cost-effective maintenance technique, this paper presents a novel research using smart sensor systems for machine fault diagnosis. In this paper, a smart sensors system is developed which acquires three types of signals involving vibration, current, and flux from induction motors. And then, support vector machine, linear discriminant analysis, k-nearest neighbors, and random forests algorithm are employed as classifiers for fault diagnosis. The parameters of these classifiers are optimized by using cross-validation method. The experimental results show that smart sensor system has the similar performance for applying in intelligent machine fault diagnosis with reduced product cost. Developed smart sensors have feasibility to apply for intelligent fault diagnosis.	['machin', 'fault', 'diagnosi', 'tradit', 'mainten', 'problem', 'past', 'mainten', 'tradit', 'sensor', 'money', 'cost', 'limit', 'wide', 'applic', 'industri', 'develop', 'cost', 'effect', 'mainten', 'techniqu', 'paper', 'present', 'novel', 'research', 'smart', 'sensor', 'system', 'machin', 'fault', 'diagnosi', 'paper', 'smart', 'sensor', 'system', 'develop', 'acquir', 'three', 'type', 'signal', 'involv', 'vibrat', 'current', 'flux', 'induct', 'motor', 'svm', 'linear', 'discrimin', 'analysi', 'knn', 'random', 'forest', 'algorithm', 'employ', 'classifi', 'fault', 'diagnosi', 'paramet', 'classifi', 'optim', 'cross', 'valid', 'method', 'experiment', 'result', 'smart', 'sensor', 'system', 'similar', 'perform', 'appli', 'intellig', 'machin', 'fault', 'diagnosi', 'reduc', 'product', 'cost', 'develop', 'smart', 'sensor', 'feasibl', 'appli', 'intellig', 'fault', 'diagnosi']

Tableau 4.2: Nettoyage de la description (Source: Auteurs)

[7] WordNet® est une grande base de données lexicale de l'anglais. Les noms, verbes, adjectifs et adverbes sont regroupés en ensembles de synonymes cognitifs (synsets), chacun exprimant un concept distinct. (PRINCETON UNIVERSITY, 2010)

4.3.4 Transformation

Suite à la revue de la littérature, nous avons vu qu'il n'y avait pas de méthode préférée pour déterminer la méthode matrice termes-documents. Après une première analyse manuelle, nous avons pu voir que certains domaines allaient être plus présents que d'autres. Par exemple, il y a beaucoup plus de documents qui ont l'air d'être reliés à la médecine qu'au transport. Au vu de cette disproportion, il a été décidé d'écarter la méthode TF-IDF qui sous-estimerait les termes liés à la médecine étant donné qu'ils seraient énormément présents comparés aux termes représentant le transport. La méthode Binaire a été préférée à la méthode de simple fréquence pour la simple raison que si un terme n'apparait que dans peu de documents c'est qu'il est peu probable qu'il soit représentatif d'un domaine. Et à l'inverse, s'il apparait dans beaucoup documents, c'est qu'il est plutôt lié à notre thème de collecte : le machine learning.

Au vu de la dimension de cette matrice, il a été décidé de la réduire comme cela est fait dans tous les travaux présentés dans la revue de la littérature. Pour cela, nous avons réalisé un tri sur les tokens afin de déterminer ceux qui étaient liés à un domaine. Cela a permis de réduire de manière considérable la matrice étant donné que chacun de nos articles applicatifs étaient composés d'au moins 2 domaines à savoir le ML et le domaine d'applications. Pour réaliser le tri, nous avons décidé de ne garder que les tokens qui apparaissaient dans au moins 10 documents. Nous considérions qu'en dessous de cette valeur, nous ne pouvions pas dire que le token était représentatif d'un domaine. Nous avons également supprimé les tokens qui apparaissaient dans plus de 93 documents. Ce choix s'est fait sur base d'une analyse, où nous avons vu que les tokens liés à un domaine ne commençaient à apparaitre qu'en dessous de cette valeur. Une analyse manuelle du reste des tokens s'est alors déroulée afin de déterminer si ceux-ci étaient liés à un domaine et ne garder que ceux-ci. Le Tableau 4.3 ci-dessous montre un exemple des tokens qu'il nous reste après le tri pour un document.

Tokens	Tokens Domaines
['credit', 'score', 'model', 'wide', 'studi', 'area', 'statist', 'machin', 'learn', 'artifici', 'intellig', 'mani', 'novel', 'approach', 'artifici', 'neural', 'network', 'ann', 'rough', 'set', 'decis', 'tree', 'propos', 'increas', 'accuraci', 'credit', 'score', 'model', 'improv', 'accuraci', 'fraction', 'percent', 'translat', 'signific', 'save', 'sophist', 'model', 'propos', 'improv', 'accuraci', 'credit', 'score', 'mode', 'paper', 'genet', 'program', 'build', 'credit', 'score', 'model', 'two', 'numer', 'exampl', 'employ', 'compar', 'error', 'rate', 'credit', 'score', 'model', 'includ', 'artifici', 'neural', 'network', 'decis', 'tree', 'rough', 'set', 'logist', 'regress', 'basi', 'result', 'conclud', 'provid', 'better', 'perform', 'model']	['credit', 'score', 'save']

Tableau 4.3: Trie tokens (Source: Auteurs)

5 Topic modeling

Afin de découvrir différents domaines d'applications présents dans le corpus, nous avons décidé d'utiliser la modélisation de sujets avec la méthode LDA. Comme nous avons pu le voir dans le point 3.1.3, cette méthode requiert de trouver un nombre de sujets. Toutefois, pour trouver ce nombre de sujets il n'existe pas de méthode exacte et les résultats restent très subjectifs à l'interprétation de l'utilisateur.

Dans le cadre de cette étude, nous avons décidé d'analyser la perplexité de différents nombres de sujets et de différentes combinaisons d'alpha et beta. Suite à cette analyse, nous avons décidé de fixé alpha à « auto » et beta à 1.5. « Si la distribution est asymétrique, une valeur **β** élevée se traduira par une distribution de mots plus spécifique. Les sujets, cependant, seront plus similaires en termes de mots contenus » (NAUSHAN, 2020). L'analyse des perplexités obtenues avec ces paramètres (Annexe 6) et notre aspect critique ont permis de déterminer que le nombre de sujets ayant le plus de sens était 12. L'annexe 7 fournit une visualisation de la distribution des sujets. Les documents ont été déclarés comme appartenant à un certain sujet si la probabilité d'appartenir à ce sujet était plus grande qu'un certain seuil déterminé manuellement. Ce seuil permet d'augmenter la probabilité de n'avoir que des articles réellement liés au sujet dominant.

L'analyse plus précise des sujets des domaines et de la répartition des domaines dans le corpus se fait à la partie 6.2.

6 Analyses et résultats

6.1 Analyse descriptive

Une fois le nettoyage des données réalisé, nous avons pu procéder à une analyse descriptive des données. L'analyse descriptive permet d'observer :

- Le nombre d'articles récoltés par chaque requête
- Le nombre d'articles récoltés par année
- Les 10 mots clés les plus utilisés
- Les 10 auteurs les plus impliqués dans les publications
- Les 10 articles les plus cités
- L'évolution des techniques de machine learning

Cela nous permet d'avoir un premier aperçu concernant les données collectées. Pour rappel, la plupart de ces observations repose à la fois sur les articles théoriques et applicatifs collectés donc 5827 articles.

6.1.1 Nombres d'articles récoltés par requête :

Pour rappel, les mots-clés utilisés pour collecter les différentes données étaient les suivants : ***machine learning, data mining, learning algorithm, decision tree, support vector machine, artificial intelligence, learning method, random forest, neural networks, supervised learning*** *et* ***unsupervised learning***. Comme la Figure 6-1 le montre, avec 2111 articles parmi les 5828 récoltés, c'est le mot *neural networks* qui a permis de récolter le plus d'articles. Les mots *random forest* et *unsupervised learning* ont quant à eux permis de récolter le moins d'articles avec respectivement 165 et 153 articles récoltés.

La Figure 6-2 permet de voir le pourcentage d'articles récoltés par un certain nombre de requêtes. La majorité des articles (61%) a été récolté via une seule requête. Il y a également 0.03% des articles qui ont été récoltés via 8 requêtes sur 11, cela représente 2 articles sur les 5828 récoltés. Ce graphique montre l'importance de procéder à un nettoyage des doublons. En effet, sans ce nettoyage, nous aurions eu plus de 3438 articles apparaissant au moins 2 fois dans nos données.

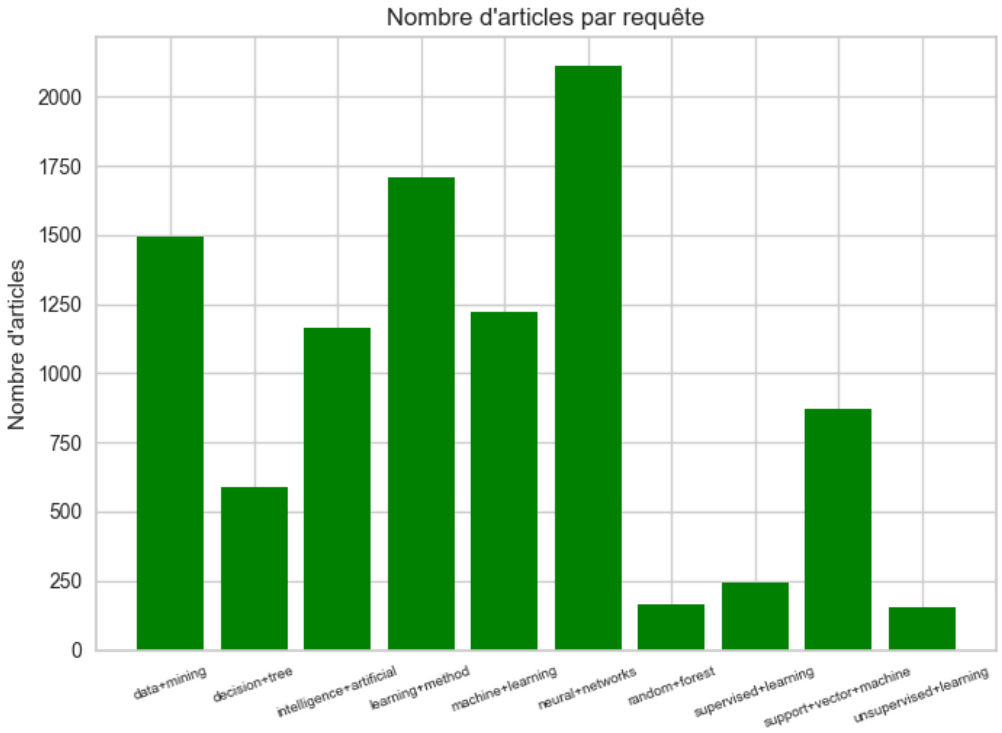


Figure 6-1: Nombre d'articles par requête (Source : Auteurs)

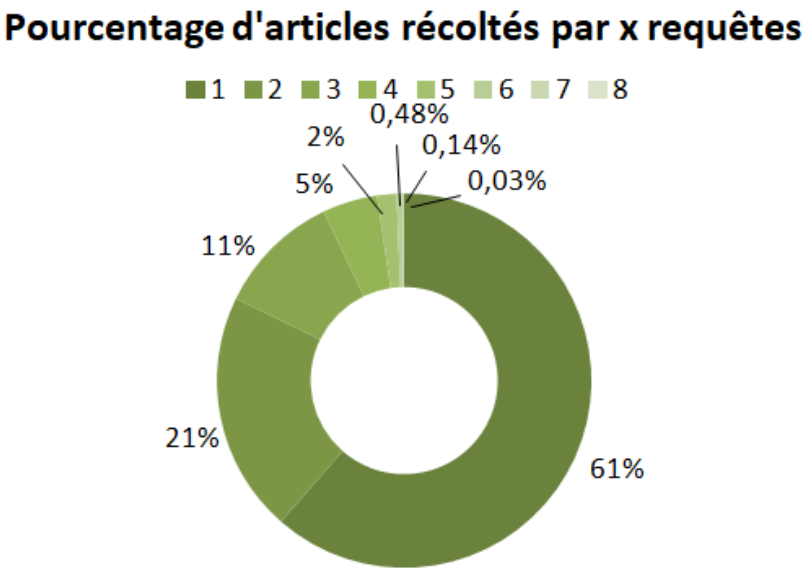


Figure 6-2: Pourcentage d'articles récoltés par x requêtes (Source: Auteurs)

6.1.2 Nombres d'articles récoltés par année

La Figure 6-3 montre l'évolution des articles récoltés en lien avec le machine learning dans le journal *Expert Systems with Applications.* Ce graphique permet d'avoir une image globale de l'importance du machine learning dans le temps. De plus, la tendance du nombre d'articles collectés suit en général la tendance du nombre d'articles publiés (Figure 6-4). Ces diagrammes permettent d'observer plusieurs tendances. Tout d'abord, il est possible de voir

une tendance à la hausse du nombre d'articles liés au machine learning jusqu'en 2002. En 2002, se situe un pic avec plus de 70% d'articles du journal qui concerne le machine learning. Ensuite, de 2004 à 2009, les publications semblent être constantes en tournant autour des 50% du nombre de publications totales. De 2010 à 2016, il y a une légère baisse du pourcentage, qui passe en dessous des 40%. Ce pourcentage remontera ensuite en 2017 pour rester aux environs de 50% jusqu'à 2020.

La présence du ML dans la littérature se distingue particulièrement à deux périodes. Une première fois aux alentours de l'année 2000 et une deuxième fois à partir de 2016. La première phase montre l'importance qu'a pris le machine learning au début des années 2000. Comme certains articles analysés dans la revue de la littérature le mentionnent, cela est notamment dû à l'augmentation des capacités de calculs des ordinateurs et à l'accès à de plus en plus de données. Pour la seconde phase, l'hypothèse de l'apparition du deep learning dans les années 2010 est émise ainsi que l'augmentation de l'utilisation d'intelligence artificielle. En effet, le deep learning est apparu vers l'année 2010. Une fois que le deep learning aura fait ses preuves, nous nous attendons à obtenir une augmentation des cas utilisant cette méthode et donc du nombre d'articles publiant les résultats de ces cas d'utilisations. Cette hypothèse sera confirmée avec une analyse plus précise sur le mot-clé *deep learning* dans le temps (Annexe5). Cela confirme également les observations de ZHANG, TAN, HAN, & ZHU (2017) et DASTILE, CELIK, & POTSANE (2020) indiquant que le deep learning et les CNN apparaissaient de plus en plus.

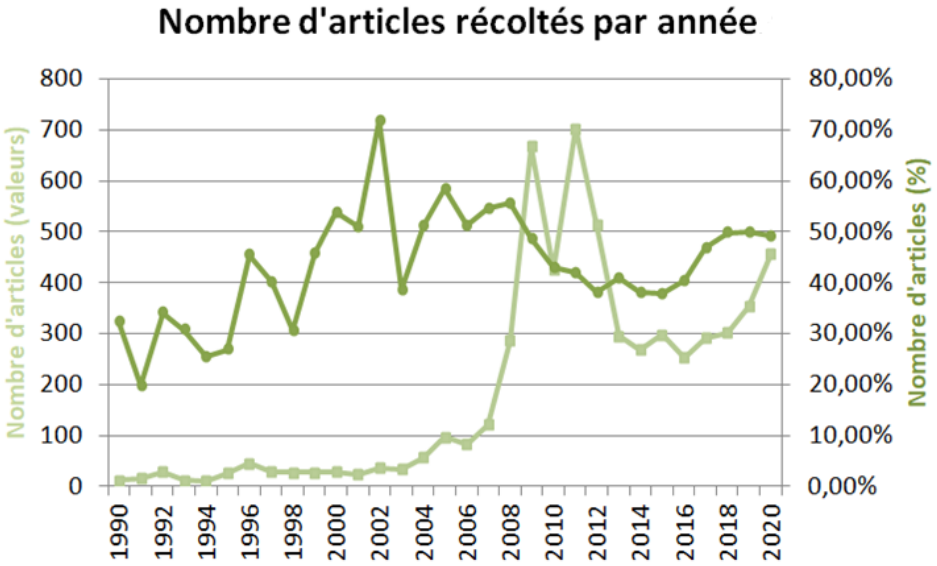


Figure 6-3: Nombre d'articles par année par rapport au nombre total d'articles publiés dans "Expert Systems with Application" (Source : Auteurs)

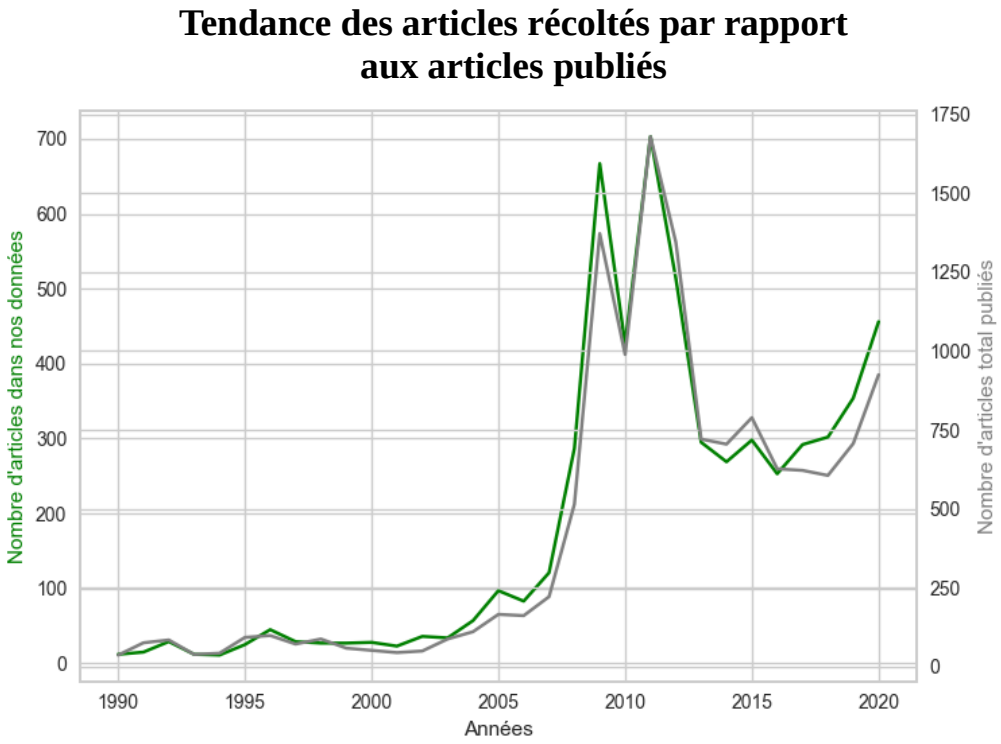


Figure 6-4: Tendance des articles récoltés par rapport aux articles publiés (Source: Auteurs)

Il est important de signaler que l'année 2020 est présente sur ce graphique uniquement à titre informatif. Aucune interprétation ne peut être faite étant donné que les données ont été collectées avant la fin de l'année 2020, il manque donc une partie des articles publiés cette année-là.

6.1.3 Top 10 des mots clés les plus utilisés

Pour analyser les mots-clés les plus utilisés, une comparaison entre les mots-clés sans nettoyage (Figure 6-5) et les mots-clés après nettoyage et génération (Figure 6-6) est nécessaire. Cela met en évidence l'impact de notre processus de nettoyage et de génération de mots-clés manquants. Globalement, le nombre d'occurrence de chaque mot-clé augmente après le nettoyage et la génération. Par exemple, le mot-clé *data mining* passe de 594 occurrences à 625. De plus, l'ordre d'importance des mots-clés est légèrement modifié. Par exemple, *expert system* apparait dans le top 10 au détriment de *deep learning*, en passant de 127 à 212 occurrences. Cela prouve encore l'apparition récente du deep learning étant donné que la génération des mots-clés s'est faite principalement pour les années 90, où le deep learning n'existait pas.

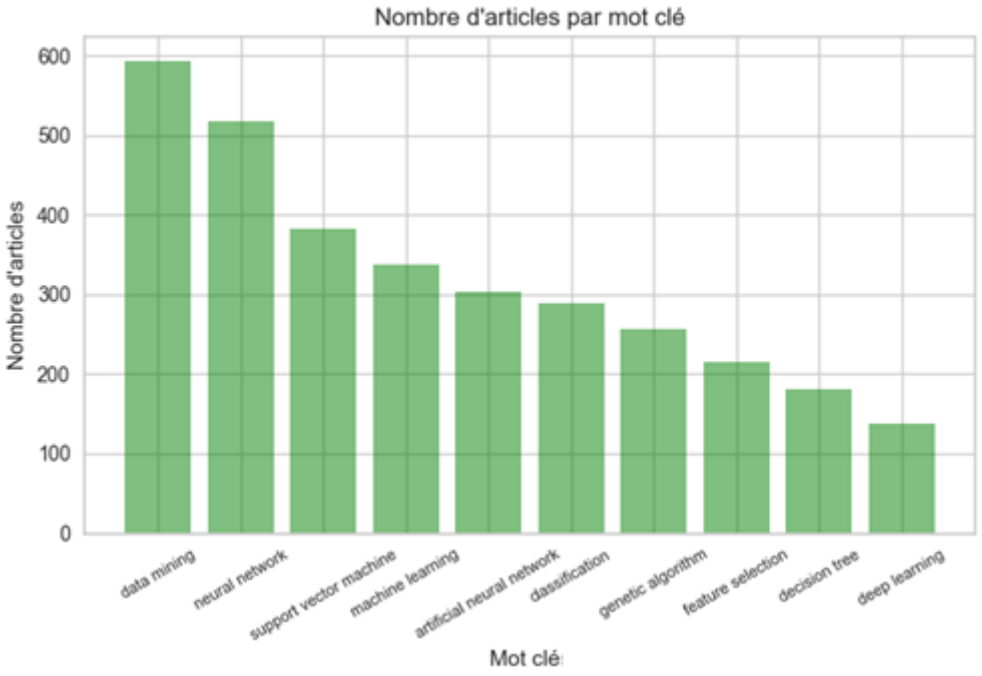


Figure 6-5: Nombre d'articles par mot clé avant nettoyage (Source : Auteurs)

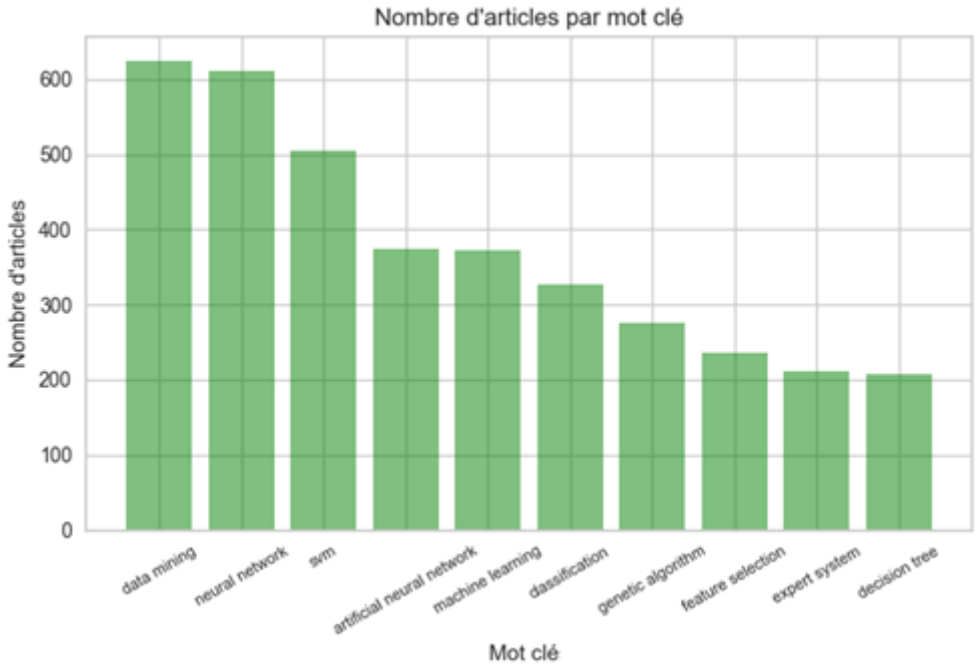


Figure 6-6: Nombre d'articles par mot-clé après nettoyage et génération mots-clés manquants (Source : Auteurs)

En se concentrant sur les mots-clés après nettoyage, nous pouvons voir que tous les mots sont liés avec le machine learning, ce qui permet de montrer la pertinence des données collectées. La Figure 6-6 suggère que les techniques de machine learning les plus utilisées sont les réseaux de neurones et les machines à vecteur de support (SVMs). Nous pouvons également dire que nos résultats confirment la conclusion de la revue de littérature étant donné que les réseaux de neurones et les SVMs sont les deux techniques les plus utilisées. Les arbres de

décisions apparaissent également dans le top 10. Cela prouve également ce qui a été vu dans la littérature à savoir que c'était une technique utilisée dans énormément de domaines dû à sa qualité d'interprétation.

6.1.4 Top 10 des auteurs les plus impliqués dans les publications du machine learning

Une analyse décrivant des 10 auteurs les plus présents dans le domaine du machine learning semble important (Figure 6-7). En effet, en cas de lecture d'articles pour obtenir plus d'informations, il semblera pertinent de lire des articles des auteurs les plus impliqués dans le machine learning. De cette manière, il est plus probable d'obtenir des informations pertinentes étant donné que ces auteurs ont une grande expérience dans le domaine. Trois auteurs se distinguent avec plus de 25 publications. Il semble donc intéressant d'obtenir plus d'informations concernant leurs domaines d'études qui seront utiles lorsque la lecture de certains articles sera nécessaire.

Le premier auteur écrivant le plus d'articles est Hong Tsug-Pei. Il enseigne à l'université de Kaohsiung à Taiwan et ces domaines d'expertises sont « le machine learning, data mining, l'intelligence informatique, les règles d'associations, l'algorithme génétique, l'extraction et l'exploitation des usages du web » (RESEARCHGATE GMBH, 2008-2021). Le deuxième est Nanni Loris, professeur à l'université de Padova et dont les compétences sont : « extraction de caractéristiques, la classification, la reconnaissance des modèles, l'intelligence artificielle, le clustering, les systèmes intelligents et les svm » (RESEARCHGATE GMBH, 2008-2021). Le troisième est Wu Qi, professeur à l'université d'Adelaide en Australie et dont les domaines d'expertises sont la vision d'ordinateur, le ML et la reconnaissance des modèles. Tous sont donc des experts dans le machine learning, ce qui est cohérent avec la recherche.

Il est également important de noter que le premier européen Dirk Van den Poel, professeur à l'université de Gent, se situe en cinquième position.

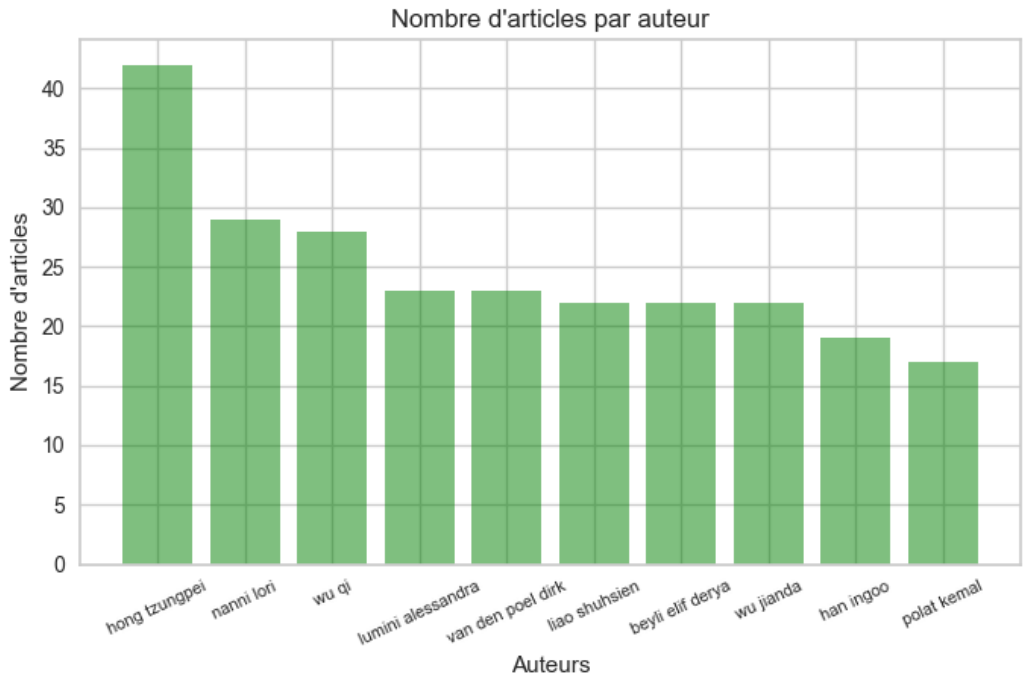


Figure 6-7: Nombre d'articles par auteur (Source : Auteurs)

6.1.5 Top 10 des articles les plus cités

Parmi les articles les plus cités (Figure 6-8), il y a deux articles parlant de SVM (n°1 avec 1055 citations et n°5 avec 626), trois articles sur le domaine de l'éducation (n°2 , n°4 et n°8 avec respectivement 865, 681 et 579 citations), deux articles dans le domaine médical et plus précisément concernant les crises d'épilepsie (n°3 avec 732 citations et n°7 avec 593), un article sur l'évaluation d'un risque concernant un pont (576 citations), un article concernant des prédictions dans le taux de change (559 citations). Cet article sera donc pertinent à lire dans le cas de prédiction liée à la finance. Le sixième article le plus cité avec 609 citations, cite différentes applications des méthodologies des systèmes d'experts de 1995 à 2004.

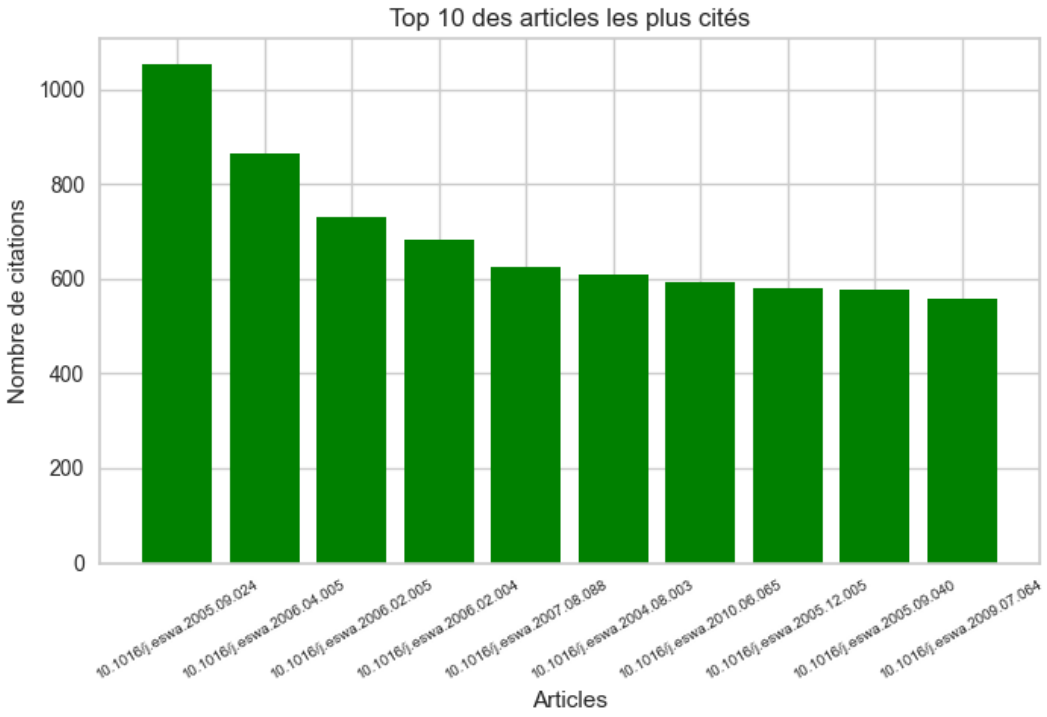


Figure 6-8: Les 10 articles les plus cités (Source: Auteurs)

6.1.6 Evolution des techniques de machine learning dans le temps en général

L'analyse des techniques de machine learning porte sur les mots clés : *knn*, *decision tree/random forest*, *regression*, *neural network*, *svm*, *deep neural network*, *bayes*, *genetic algorithm*. Ce choix a été réalisé étant donné que ce sont les techniques de ML learning les plus connues[8]. La Figure 6-9 présente l'évolution de ces techniques de machine learning dans le temps. Elle révèle que parmi tous les articles parlant des techniques de machine learning, les réseaux de neurones sont les plus sollicités. Toutefois, ils subissent une baisse depuis les années 2000. Les SVMs quant à eux ont connu une présence particulière de 2004 à 2016, mais diminuent à partir de 2016. Depuis 2015, les réseaux de neurones connaissent à nouveau une hausse mais ce sont des réseaux de neurones profonds qui sont appelés « deep neural network ». Les réseaux de neurones simples continuent à diminuer. Il y a donc une deuxième renaissance des réseaux de neurones grâce au deep learning. Au vu de ce graphique, nous observons que l'intérêt porté sur les SVMs augmente au détriment des réseaux de neurones vers l'année 2004. Il se tasse avec l'apparition du deep learning vers 2015. La vague de croissance des SVMs observée rejoint les résultats obtenus par LIN, HU, & TSAI (2011), indiquant que les SVMs sont présents à partir de 2004. Cela confirme également la forte augmentation dans les années 2010 présentée par VOYANT, et al. (2017) et MOSAVI,

[8] Présentation des techniques les plus connues du ML dans le cours de ML de Mr B.Frenay

OZTURK, & CHAU (2018). On observe que malgré que les arbres de décision/random forest ne soient jamais la technique la plus utilisée, ils restent quand même une technique avec une présence constante (entre 10 et 20%) et se situent souvent dans le top 3. Depuis les années 2004, c'est effectivement une technique de référence prisée notamment pour l'exploitation et l'interprétation de ses résultats comme nous avons pu le voir dans la revue de littérature. Nous pouvons également voir une légère diminution dans l'utilisation de l'algorithme génétique qui était une des techniques les plus utilisées avant les années 2000. Toutefois, avec l'apparition des SVMs, cette technique a tendance à être de moins en moins utilisée.

Cette évolution confirme l'analyse des mots-clés les plus utilisés (partie 6.1.3) avec tout d'abord les réseaux de neurones, puis les SVMs, les arbres de décisions/random forest avec leur utilisation constante et le deep learning avec son évolution exponentielle ces dernières années. Cette analyse confirme également la revue de la littérature, où nous avons vu que ces 3 techniques étaient les plus utilisées dans la plupart des domaines et que le deep learning était en évolution et que les utilisations et les recherches à ce sujet allaient encore augmenter.

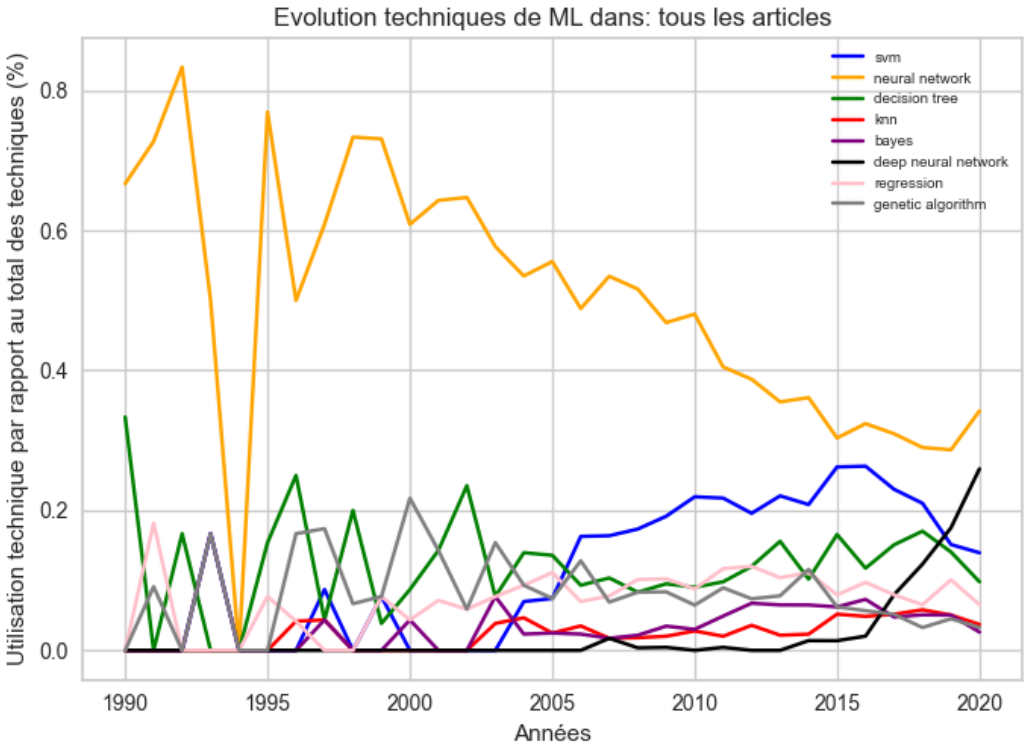


Figure 6-9: Evolution techniques de ML dans le temps (Source: Auteurs)

6.2 Résultats

6.2.1 Evolution techniques de machine learning dans les articles applicatifs

La Figure 6-10 montre l'évolution des usages des techniques de machine learning. Cette analyse a été réalisée uniquement sur les articles qui ont été considérés comme applicatifs. Cette figure confirme les tendances qui ont été expliquées dans le point 6.1.6. Cependant, en comparant avec l'évolution théorique et applicative des techniques de ML, nous pouvons voir que l'application a eu tendance à commencer uniquement dans les années 2000. Avant cette année-là, très peu d'articles parlaient de l'usage des techniques de ML, ils étaient plus orientés théorie. Une hypothèse à cela est qu'avant les années 2000, très peu d'articles scientifiques étaient publiés digitalement et l'accent était surtout mis sur les articles théoriques afin de partager les connaissances avec un plus grand nombre de personnes.

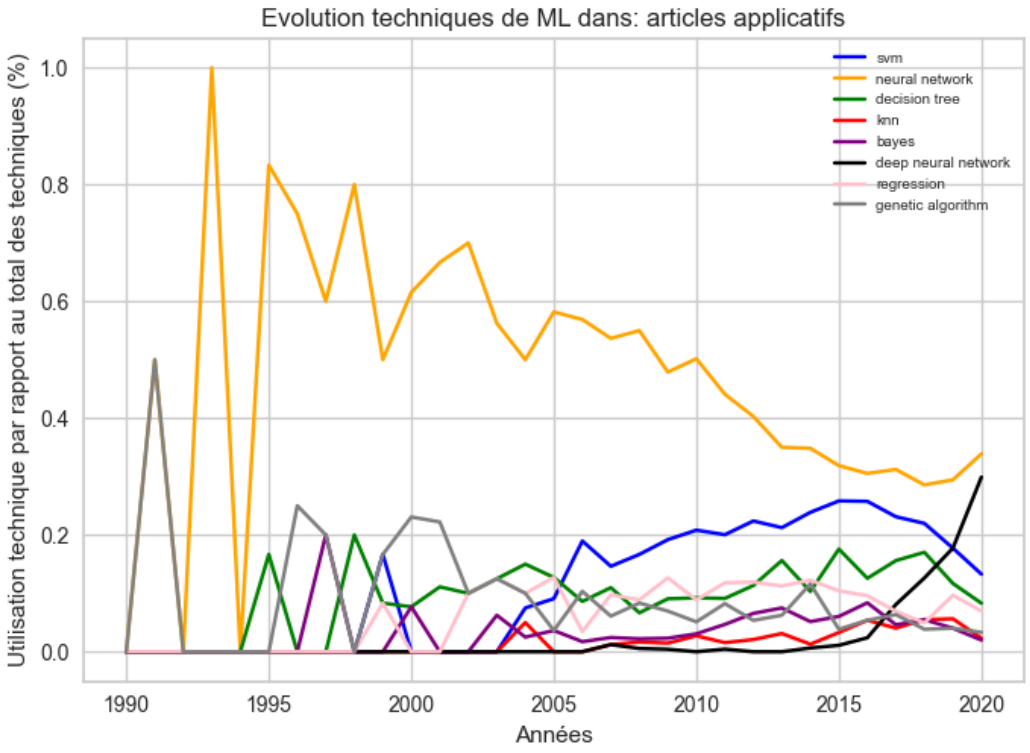


Figure 6-10: Evolution usage des techniques de ML (Source: Auteurs)

6.2.2 Nuage de mots des tokens liés aux domaines.

La Figure 6-11 est un nuage de mots représentant la fréquence des différents tokens liés aux domaines présents dans le corpus. Plus un mot est large, plus il est présent. Au vu de cette figure, nous pouvons en déduire que les domaines d'applications de la finance, le monde de l'entreprise, de la médecine sont fortement présents dans notre corpus. Car pour la finance, les mots ***financi, market, trend*** sont fortement présents. Pour l'entreprise, nous avons les mots

custom, busi, demand, compani, qui ressortent. Enfin, pour le domaine médical, il y a les mots ***patient, medic, disease, clinic***. Cela confirme également la revue de la littérature où nous avons pu voir que ces 3 domaines faisaient partie des domaines utilisant énormément le machine learning. Cela rejoint également l'analyse des articles les plus cités où nous avions vu qu'il y avait des articles du domaine médical et de la finance.

Nous pouvons également nous attendre à avoir des articles liés au transport ou à des mouvements étant donné que nous voyons les mots ***speed, move, transport, vehicle***. De même, avec les mots ***sentiment, review, content, social, platform,*** nous nous attendons à avoir un domaine lié à l'analyse du contenu sur internet.

Figure 6-11: Nuage de mots tokens (Source: Auteurs)

6.2.3 Analyse des domaines

La modélisation de sujet a permis de faire ressortir 12 sujets du corpus. Parmi ceux-ci, 11 semblent avoir une signification. Le Tableau 6.1 présente tous les sujets ainsi que les 10 mots les plus représentatifs. La labellisation des sujets a été réalisée sur base de la connaissance.

Le premier domaine se rapporte à l'analyse des ***signaux physiologiques***. Nous pouvons voir que ce sujet comprend également les mots liés aux crises d'épilepsie. Cela est dû au fait que les électroencéphalogrammes sont très souvent utilisés dans les recherches analysant le cerveau des personnes épileptiques. La détection de ce sujet a du sens au vu des articles les plus cités. En effet, nous avions vu que 2 des articles les plus cités étaient liés aux

crises d'épilepsie. Cela montrait donc que c'était un domaine important dans l'utilisation du machine learning.

Le deuxième sujet concerne la ***biologie***. Cette étiquette a été donnée au vu des mots ***protein, acid, cell, molecular, genom, stom, atom, protocol***.

Le troisième sujet se rapporte à la ***culture***. Cette étiquette a été donnée au vu des différents mots représentant l'éducation, la musique mais également via l'analyse des documents représentatifs de ce sujet. Trois articles sur l'éducation étaient présents dans le top 10 des articles les plus cités, il fait donc sens que la modélisation de sujet ait trouvé un sujet lié.

Le quatrième sujet fait référence à ***l'analyse de texte***. Au vu des mots-clés (***social, twitter***) et de l'analyse des documents liés, une majorité de cette analyse de texte a lieu sur des textes provenant du web. Ce sujet confirme l'hypothèse faite dans l'analyse du nuage de tokens qui déterminait que l'analyse du contenu sur internet ferait partie des sujets.

Le cinquième sujet fait référence à la ***finance***, et plus généralement au prix des actions et de leurs tendances. Au vu des mots clés ***trend*** et ***volatif***, nous nous attendons à avoir plusieurs documents liés à la prédiction de changement de prix des actions et donc à l'utilisation des techniques de machine learning qui permettent la prédiction. Ce sujet confirme également une hypothèse faite via l'analyse du nuage de tokens. Ce sujet fait également sens car nous avions vu qu'un article concernant la finance faisait partie du top 10 des articles les plus cités.

Pour le sixième sujet, nous n'avons pas su donné d'étiquette. Au vu des différents mots-clés, nous pourrions pensés à une gestion de l'écologie au sein des entreprises. Cependant, l'analyse des documents récoltés à révéler avoir trop de documents qui n'avaient pas de lien entre eux et avec ce sujet. C'est pourquoi, nous n'avons pas réussi à donner une signification à ce sujet.

Le septième sujet était surtout lié à des mots et documents liés à la gestion et aux contrôles d'outils, de machines et d'engins, c'est pourquoi nous avons décidé de lui donner le nom général ***manufacture***.

Le huitième sujet représente la ***médecine*** en générale, comme nous l'avions prédit dans le point précédent. Au vu des mots-clés ***disease, treatment*** et ***diagnos***, nous pouvons

nous attendre à avoir plusieurs articles concernant la classification de maladie et donc des algorithmes de classification.

Le neuvième sujet est lié aux ***cancers et maladies du cœur***. Ces deux maladies ont certainement été traitées comme un sujet étant donné qu'énormément d'articles scientifiques sont reliés à la recherche dans ces maladies.

Le dixième sujet confirme également une hypothèse faite dans l'analyse du nuage de mots des tokens. Ce sujet est lié à la ***gestion d'entreprise*** en général. Cela englobe aussi bien la gestion des clients, la compétition entre entreprises ou la finance d'entreprise. Via la revue de la littérature, nous avions vu que la gestion d'entreprise utilisait énormément le machine learning.

L'onzième sujet fait référence à l'intrusion ***et à la détection de défaut***, donc de manière générale, à quelque chose qui ne passe pas comme prévu et qui n'est pas souhaité. Comme nous pouvons le voir avec les mots ***protect, biometr*** et ***privaci***, nous nous attendons à avoir dans ce sujet également des articles qui parlent de la sécurité pour ne pas devoir faire face à ces intrusions et défauts.

Le dernier sujet fait quant à lui référence à ***l'analyse de mouvement***. Les mots-clés ***vehicle*** et ***transport***, nous font penser qu'une partie de ces articles est liée au transport en général. Ce sujet n'est pas uniquement lié au transport car des mouvements peuvent également être analysés dans la voix, via les vidéos surveillances, etc.

Sujets→ Mots clés↓	Signal	Biologie	culture	analyse textes	finance	Colonne1	manufacture	médical	cancer-cœur	entreprise	intrusion - défaut	mouvement
1	eeg	protein	student	social	financi	load	monitor	patient	cancer	custom	attack	vehicle
2	electroencephalogram	scid	educ	content	market	plant	industri	disease	breast	industri	protect	video
3	epilept	cell	cours	sentiment	stock	water	temperatur	medic	ecg	market	intrus	speed
4	epilepsi	molecular	teacher	opinion	price	treatment	speed	clinic	tumor	competit	biometr	transport
5	spectral	genom	teach	corpus	trade	delay	surface	health	electrocardiogram	busi	privaci	trajectori
6	patient	www	question	sentenc	return	voltage	defect	treatment	beat	compani	histogram	mobil
7	band	window	music	medium	trend	conflict	yield	care	arrhythmia	demand	malwar	monitor
8	channel	stom	speaker	textual	ratio	wastewat	manufactur	diagnos	benign	consum	threat	motion
9	eye	atom	tutor	twitter	economi	channel	electr	heart	morpholog	manufactur	server	move
10	healthi	protocol	school	share	volatif	pollut	respons	hospit	ultrasound	credit	color	track

Tableau 6.1: Top 10 tokens dans les different domains (Source: Auteurs)

6.2.4 Répartition des articles dans les domaines

La Figure 6-12 ci-dessous représente le nombre de publications associées à chaque sujet. Nous pouvons voir que le sujet avec le plus de publications est la gestion d'entreprise avec

720 articles sur les 3687 articles applicatifs collectés. Il y a ensuite le domaine de la finance avec 461 articles, la manufacture avec 430 articles, les mouvements avec 412 articles et le médical avec 411 articles. Au total, 3240 articles sur 3687 ont été reliés à au moins un sujet. Ce graphique permet également de montrer la grande disproportion entre les différents sujets avec 5 sujets avec moins de 150 publications et 6 avec plus de 300 publications.

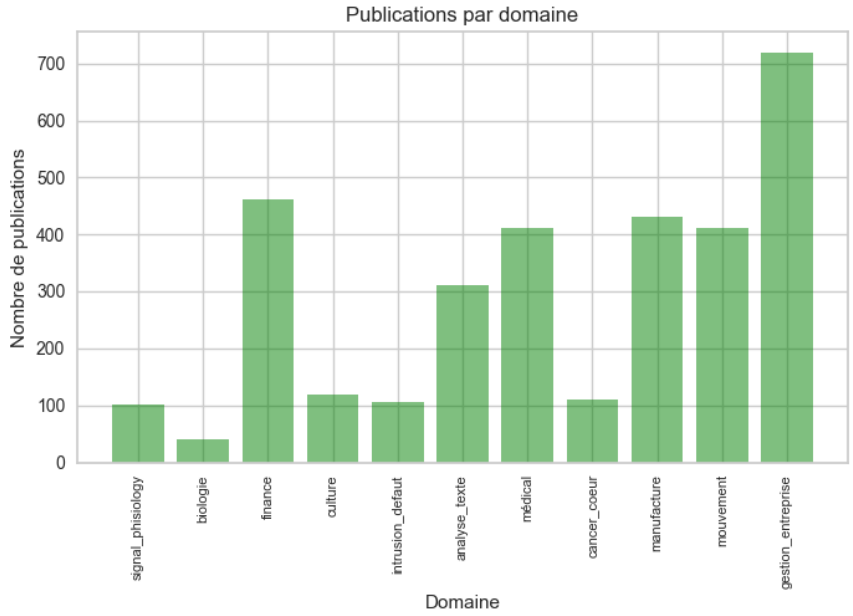


Figure 6-12: Nombre de publication par domaine (Source: Auteurs)

6.2.5 Analyse de l'évolution des publications dans les domaines

Nous avons réalisé une analyse de l'évolution des publications dans chaque sujet (Annexe 8) afin d'analyser si un évènement externe avait influencé l'augmentation ou la diminution des publications dans un des domaines. Nous avons pu remarquer plusieurs tendances.

Tout d'abord, nous pouvons voir que pour l'analyse de texte (Figure 6-13), le pourcentage varie fortement avant l'année 2004. A partir de 2004, la variation de pourcentage est beaucoup plus faible et nous pouvons voir une tendance à la hausse. Il y a même une large augmentation du pourcentage d'articles applicatifs liés à l'analyse de texte à partir de 2011 passant d'environ 5% à 15% en 2013 et atteignant plus de 30 publications en 2020. L'augmentation à partir de 2004 est notamment due à l'apparition de réseaux sociaux (Twitter plus précisément) et à une augmentation du nombre de site web, forums, blogs où l'homme peut faire part de son avis et de sa façon de penser. L'analyse de l'opinion est très utilisée pour différentes marques afin de réussir à satisfaire le client et améliorer les rendements de l'entreprise. Comme nous l'avions vu dans la revue de la littérature, avec l'utilisation du NLP.

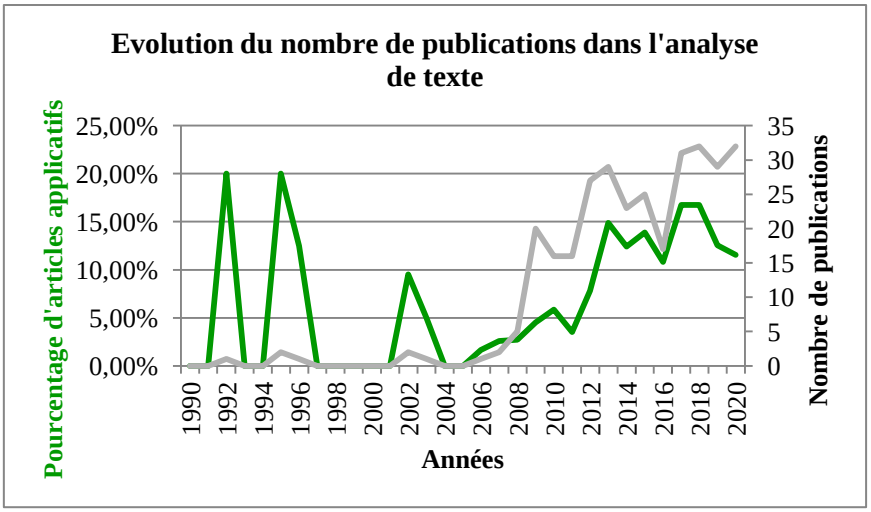


Figure 6-13: Evolution publications dans l'analyse de texte (Source: Auteurs)

Une autre observation se situe au niveau de la détection de cancer et maladie cardiaque. Nous pouvons voir sur la Figure 6-14 que les publications apparaissent à partir de l'année 2004. Cela est en partie dû à l'apparition des SVMs, qui vont se révéler être une méthode très utilisée dans la classification et la prédiction des cancers.

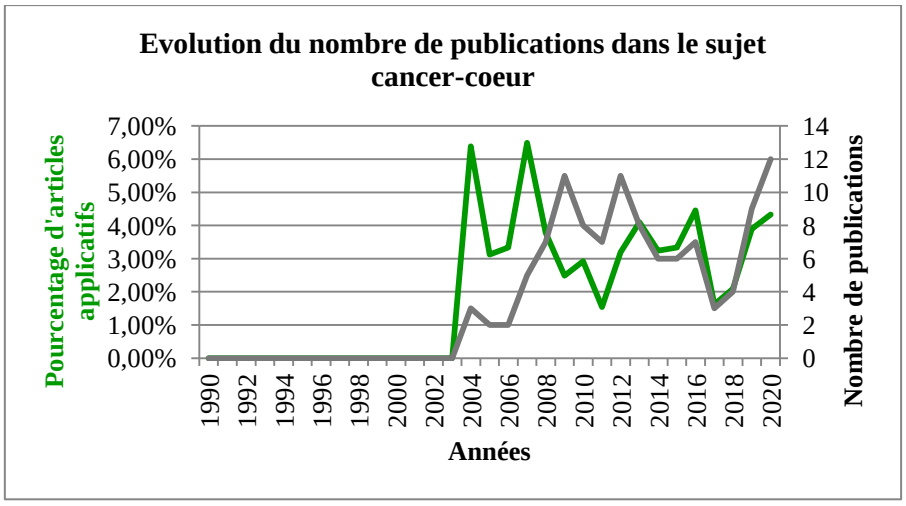


Figure 6-14: Evolution publications dans les cancers et les maladies du coeur (Source: Auteurs)

Nous pouvons également voir une diminution dans le pourcentage de publication des articles liés à la gestion d'entreprise même si les publications dans ce domaine restent très élevées (Figure 6-15). Cela peut s'expliquer par l'amélioration des techniques de machine learning qui implique que celles-ci peuvent s'appliquer à d'autres domaines que la gestion d'entreprises. Il y a donc une augmentation des recherches dans d'autres domaines ce qui fait diminuer les recherches dans ce domaine.

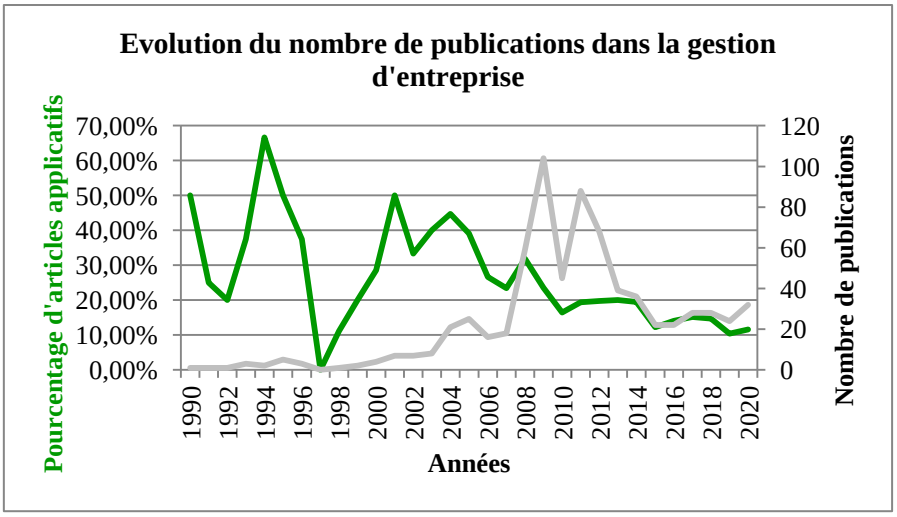


Figure 6-15: Evolution publications dans la gestion d'entreprise (Source: Auteurs)

Une autre observation concerne l'intrusion et la détection de défauts (Figure 6-16). Nous pouvons voir que depuis 2014, il y a une tendance à l'augmentation du nombre de publications. L'hypothèse émise est l'augmentation des objets intelligents qui permettent de détecter les défauts ainsi que une importance accrue à la sécurité sur le net.

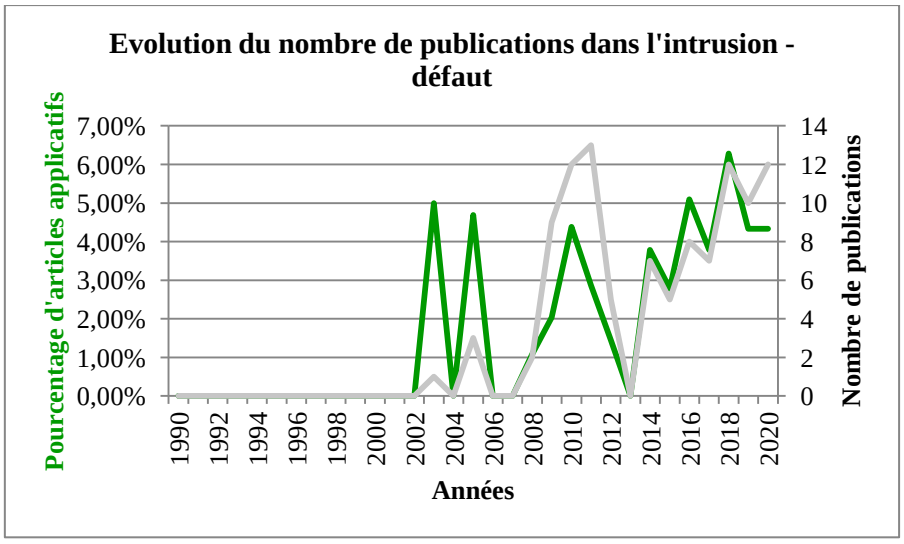


Figure 6-16: Evolution publications intrusion-défaut (Source: Auteurs)

Pour finir, en ce qui concerne l'analyse des mouvements, nous pouvons voir une augmentation ces 5 dernières années passant de 20 articles à plus de 45 (Figure 6-17). Celle-ci peut s'expliquer par l'apparition du deep learning qui permet d'analyser des mouvements via des capteurs ou sur une caméra.

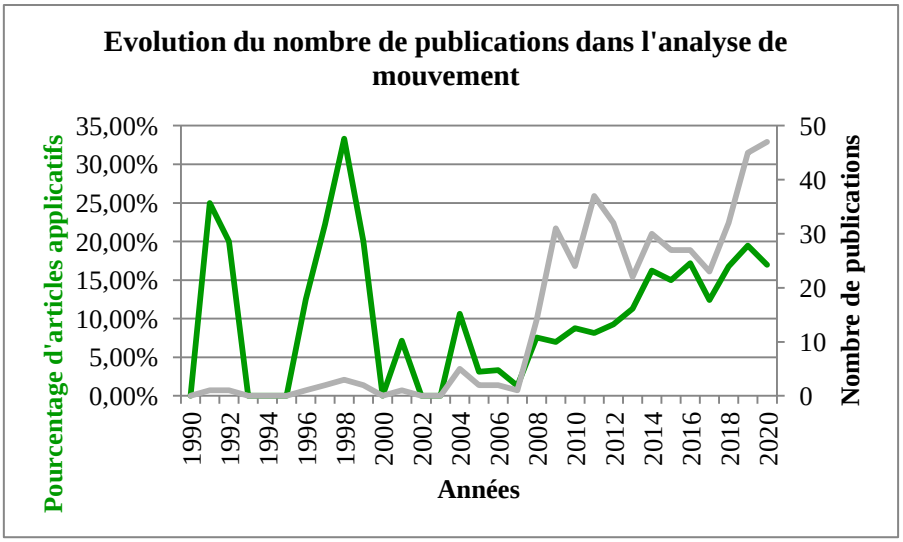


Figure 6-17: Evolution publications dans l'analyse de movement (Source: Auteurs)

6.2.6 Analyse de l'évolution des techniques de ML dans différents domaines

L'analyse des techniques de ML s'est faite sur l'apparition de ces techniques dans la description de l'article. Etant donné qu'il y a très peu d'articles applicatifs avant les années 2000, il est difficile de donner une explication de la situation à cette période. L'analyse est donc axée après les années 2000. Les techniques analysées sont : svm, réseaux de neurones, arbre de décision/random forest, knn, bayes, deep learning, régression et algorithme génétique.

L'analyse se fait sur les descriptions et non sur les mots-clés. Le choix ne s'est pas porté sur les mots-clés étant donné que certains auteurs ne mentionnaient pas la technique utilisée dans les mots-clés mais dans la description. De plus, comme il avait été mentionné par DELEN & CROSSLAN (2008), la description reprend les informations des mots-clés. Toutefois, certains auteurs ne mentionnent pas la technique utilisée ni dans les mots-clés, ni dans la description. Il y a donc une certaine perte d'informations.

Afin d'obtenir une bonne comparaison entre l'utilisation des techniques, nous utilisons la mesure suivante : le nombre de publications contenant la technique de ML mentionnée divisé par le nombre total de publications qui contiennent au moins une des techniques analysées.

Si nous regardons les sujets de la ***médecine*** (Figure 6-19) ***et du cancer – cœur*** (Figure 6-18), nous pouvons voir qu'il y a une augmentation des réseaux de neurones profonds ces dernières années. Cela généralise l'idée de ZHANG, TAN, HAN, & ZHU (2017), qui indiquait que le deep learning allait devenir une méthode très utilisée pour la découverte de médicaments. Nous voyons donc qu'elle devient une méthode très utilisée dans l'ensemble du

domaine médical. Nous pouvons également voir l'importance que les SVMs ont dans la médecine depuis leurs apparitions. Cela se voit particulièrement dans le sujet des cancers et des maladies cardiaques où les taux sont plusieurs fois supérieurs à 40%. Une autre observation est l'utilisation constante des arbres de décisions et des forêts aléatoires. Cela reflète ce qui était dit par KONONENKO (2001) quant à l'importance d'avoir un pouvoir explicatif. La régression a l'air d'être également une technique à utilisation constante avec un pourcentage tournant autour de 10%. Ces 2 figures et l'analyse du nombre de publications dans la médecine, confirment les propos de KONONENKO (2001) qui annonçait que la médecine allait avoir du mal à utiliser les techniques de ML au début des années 2000.

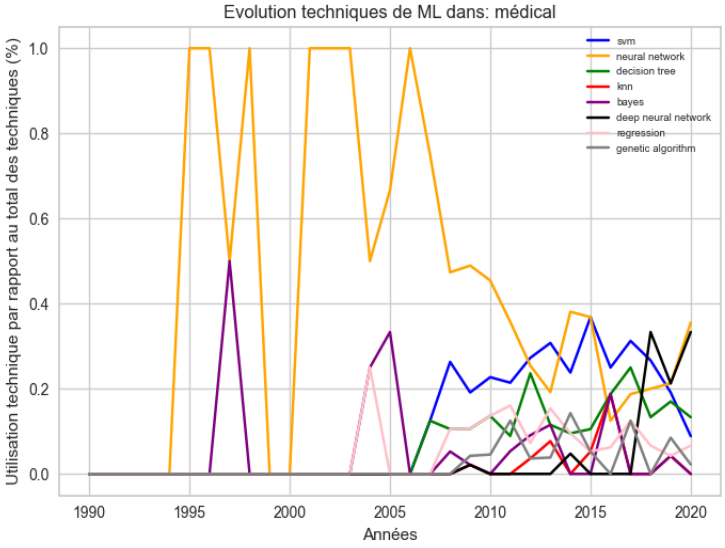


Figure 6-19: Evolution ML dans médical (Source: Auteurs)

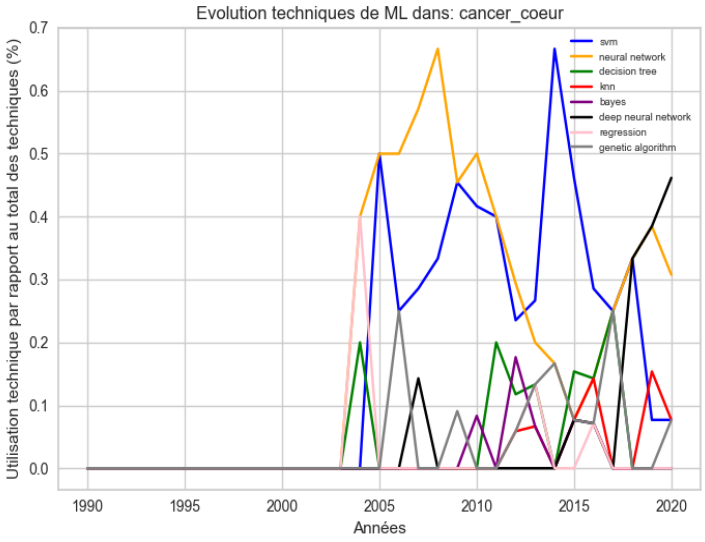


Figure 6-18: Evolution ML dans cancer – coeur (Source: Auteurs)

Pour l'analyse du domaine de ***la finance*** (Figure 6-20), nous pouvons voir qu'il y a une tendance à utiliser de plus en plus les réseaux de neurones profonds. Cela rejoint l'analyse faite par SADATRASOUL, GHOLAMIAN, SIAMI, & HAJIMOHAMMADI (2013), qui démontrait que les réseaux de neurones profonds étaient encore peu utilisés en 2013 mais avec une bonne qualité de classification. Nous pouvons également voir une utilisation assez constante tournant autour de 20% pour la régression. Cela confirme les propos de QU, QUAN, LEI, & SHI (2019) montrant que la régression logistique était très utilisée pour la prédiction de faillite. Les arbres de décisions ont une utilisation constante comme dans plusieurs domaines due à la capacité d'interprétation. Nous pouvons également voir que l'algorithme génétique a un faible pourcentage mais celui-ci semble tourner aux alentours de 10% et avait plus d'importance avant l'année 2005. Comme le mentionnaient LIN, HU, & TSAI (2011), « l'algorithme génétique est largement utilisé pour optimiser les paramètres

pour entrainer les SVMs et réseaux de neurones », d'où son utilisation constante. Globalement, les techniques les plus utilisées sont les réseaux de neurones, les SVMs, la régression et les arbres de décisions.

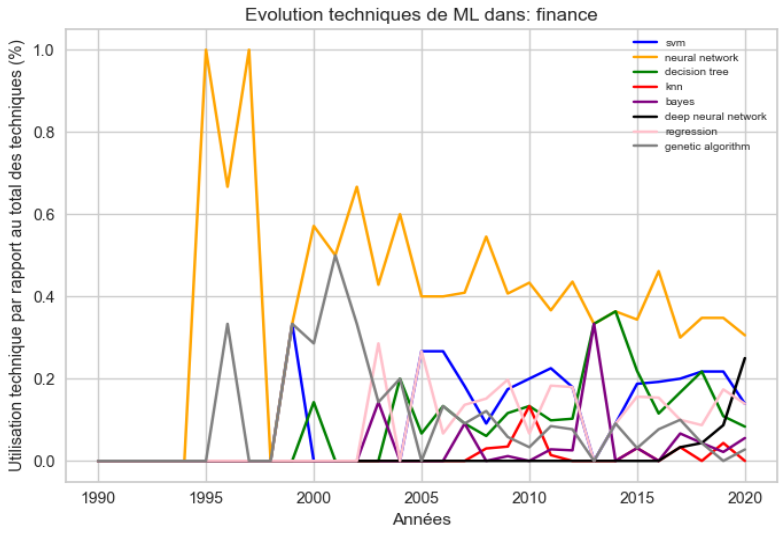


Figure 6-20: Evolution ML dans la finance (Source: Auteurs)

Pour le secteur de ***l'industrie*** (Figure 6-21), nous pouvons confirmer ce qui a été vu dans la revue de la littérature par rapport aux SVMs. En effet, ceux-ci se hissent dans le top 3 des techniques les plus utilisées dès le début de leurs apparitions. Contrairement à GE, SONG, DING & HUANG (2017), dont le classement des techniques les plus utilisées entre 2000 et 2015 était : les réseaux de neurones, régression et SVMs et puis arbres de décisions ; nos analyses démontrent que ce sont plutôt les réseaux de neurones, les SVMs, les régressions et puis les arbres de décisions. Contrairement aux autres domaines, il n'y a pas de forte augmentation dans l'utilisation du deep learning.

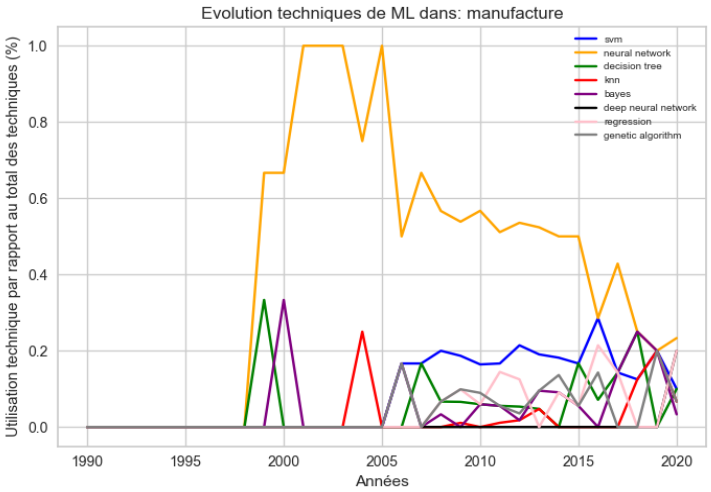


Figure 6-21: Evolution ML manufacture (Source: Auteurs)

Pour ***l'analyse de texte*** (Figure 6-22), il y a une grande importance des SVMs depuis leur apparition. Comme mentionné plus haut, l'analyse de texte a connu une croissance notamment avec l'apparition de réseaux sociaux comme Twitter. Enormément d'analyses sur Twitter sont liées à l'analyse des opinions des clients, cela demande donc une classification. Les SVMs se révèlent être un très bon algorithme de classification. Les SVMs sont également très utilisés pour la classification de mail comme spam qui requiert de l'analyse de texte. Nous pouvons voir également que la classification bayésienne est très utilisée avec utilisation tournant entre 10 et 20% avant 2017. Toutefois, l'apparition du deep learning a fait diminuer l'usage de la classification bayésienne ainsi que des SVMs.

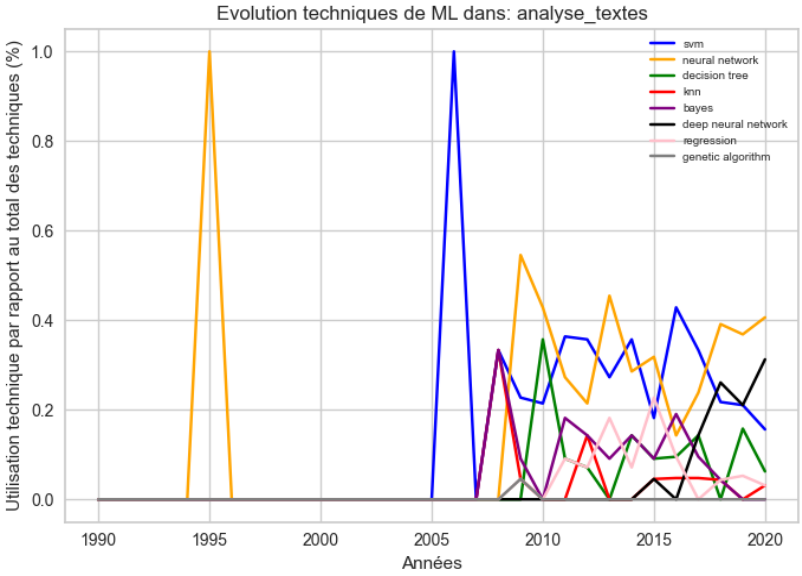


Figure 6-22: Evaluation ML dans l'analyse de texte (Source: Auteurs)

Pour la ***détection d'intrusion et de défaut*** (Figure 6-23), nous pouvons observer que les algorithmes de classification des réseaux de neurones, SVMs et arbres de décisions semblent se valoir étant donné qu'il n'y en a pas un qui semble se distinguer. GE, SONG, DING, & HUANG (2017) avaient analysé la présence des techniques de ML pour la détection de défaut au sein des industries. Les analyses avaient démontré que les SVMs et les réseaux de neurones étaient deux techniques fortement utilisées pour cette détection et qu'il y avait peu de différence entre les usages. Notre analyse rejoint donc cette étude. Le deep learning semble avoir un effet sur l'usage des techniques de ML. Cependant, l'effet est moindre comparé à celui qu'il a pu avoir dans d'autres secteurs.

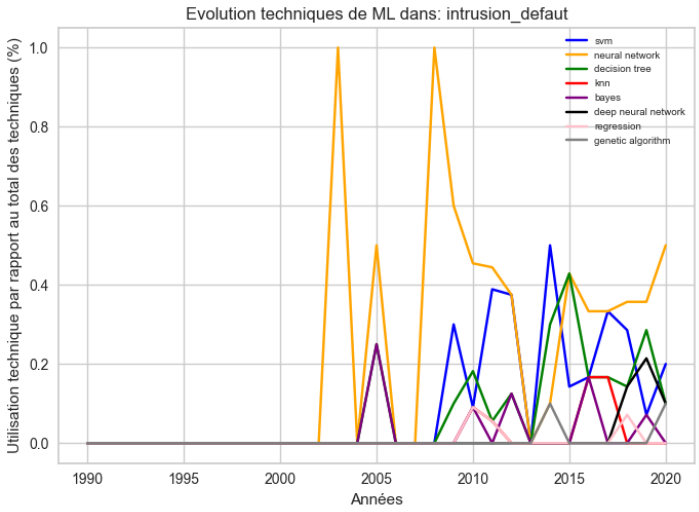


Figure 6-23: Evolution ML intrusion – defaults (Source: Auteurs)

Pour tous les articles reliés au monde de l'***entreprise*** en général (Figure 6-24), nous pouvons voir une forte utilisation des réseaux de neurones qui a toutefois tendance à diminuer aux profits d'autres techniques. Ceci s'explique par le fait que cette technique peut être utilisée dans la classification mais également dans la prédiction. Des exemples de classifications dans la gestion d'entreprises concernent le risque ou le type de client. Des exemples de prédictions sont la prédiction de la demande, des ventes ou la perte de clients. Les arbres de décisions ont une utilisation constante. Cela peut s'expliquer notamment par le fait qu'une grande partie de la gestion d'entreprise est liée à un contact humain avec les clients ou fournisseurs, qui demandent à avoir des explications. Comme il était mentionné par NGAI, WIU & CHAU (2009), l'interprétation des résultats est importante pour savoir comprendre le client dans le marketing. Comme dans d'autres domaines, la technique des SVMs est utilisée de manière constante dans le temps depuis son apparition pour ses qualités de classification. Il est important de noter que les articles récoltés pour cette analyse

comprennent des articles récoltés dans le domaine de la finance et également dans le domaine de la manufacture. Il y a donc des liens entre ces analyses. Nous pouvons également voir une utilisation constante aux alentours de 15% de la régression. Comme l'indiquaient CARBONNEAU, LAFRAMBOISE, & VAHIDOV (2008), cette méthode est très utile pour la prédiction de la demande. Cela fait donc sens d'avoir une utilisation constante car c'est une activité à laquelle les entreprises doivent toujours faire face.

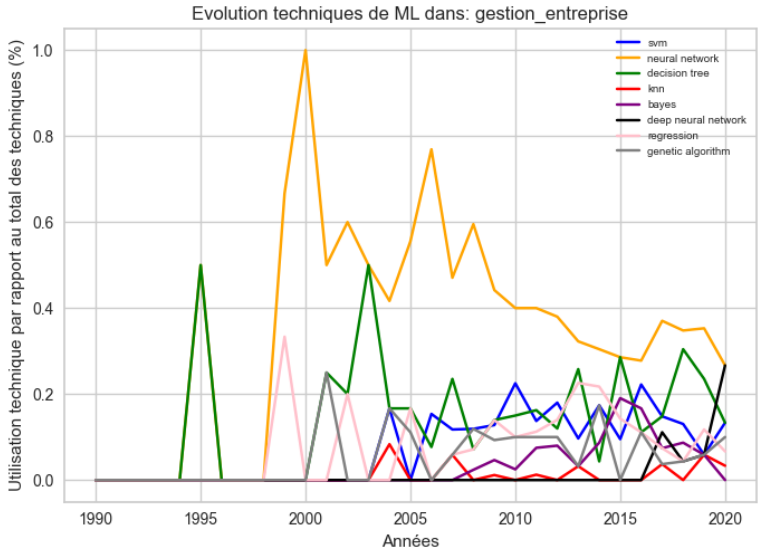


Figure 6-24: Evolution ML dans la gestion d'entreprise (Source: Auteurs)

En ce qui concerne les observations dans les ***signaux physiologiques*** (Figure 6-25), nous pouvons voir que les réseaux de neurones et les SVMs sont les deux techniques les plus utilisées dans le temps. Celles-ci sont particulièrement utiles pour la classification et la détection de crises d'épilepsie. Elles sont également utilisées pour détecter des maladies comme la schizophrénie ou la détection de la fatigue. Contrairement aux autres domaines, l'utilisation des arbres décisions semblent moins importante et non constante. Par contre, l'algorithme de clustering knn semble être plus utilisé.

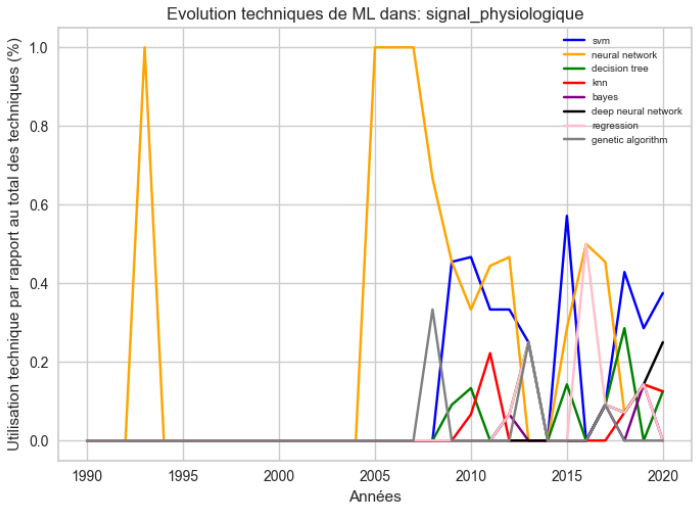


Figure 6-25: Evolution ML dans signaux physiologiques (Source: Auteurs)

Dans ***l'analyse des mouvements*** (Figure 6-26), nous pouvons observer une importance accordée aux réseaux de neurones profonds. Cela est notamment dû au fait que l'analyse de ces mouvements se fait en grande partie via des analyses vidéos comme par exemple des vidéos de la rue qui analysent le trafic routier ou des vidéos analysant le langage des signes. Les SVMs sont également fortement sollicités pour une classification des mouvements. Par exemple, pour identifier les mouvements des piétons via des capteurs pour une voiture autonome ou encore identifier la chute d'une personne en fonction du son émis lors de l'impact au sol. Nous pouvons voir que toutes ces analyses de mouvements se font via des capteurs ou analyses vidéos.

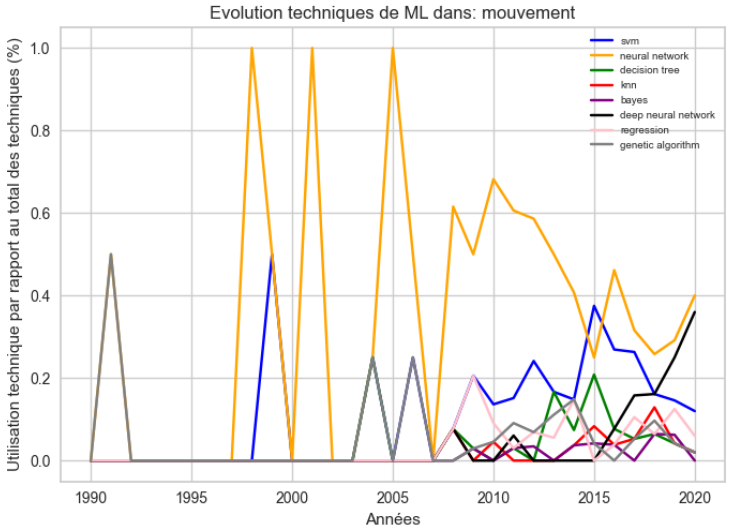


Figure 6-26: Evolution ML dans l'analyse de movement (Source: Auteurs)

Dans le domaine de la ***culture*** (Figure 6-27), qui comprend principalement des articles sur l'éducation et la musique, nous pouvons voir qu'aucune méthode ne semble réellement se distinguer. Les 3 méthodes les plus utilisées sont les réseaux de neurones, les SVMs et la régression. Ces dernières années, le deep learning prend de plus en plus d'importance comme dans la majorité des domaines. L'utilisation de la régression sert par exemple à déterminer la satisfaction des étudiants vis-à-vis de leurs cours, ou encore trier les étudiants en fonction de leurs performances académiques. Les réseaux de neurones sont utilisés dans des systèmes de recommandations de musiques ou pour détecter des étudiants qui ont un don dans les mathématiques. Les SVMs peuvent être utilisés pour reconnaitre des écritures dans différents ouvrages ou segmenter les interlocuteurs à partir des sons émis. Les réseaux de neurones profonds peuvent être utilisés pour analyser des vidéos et reconnaitre des scènes de film ou encore transcrire des musiques.

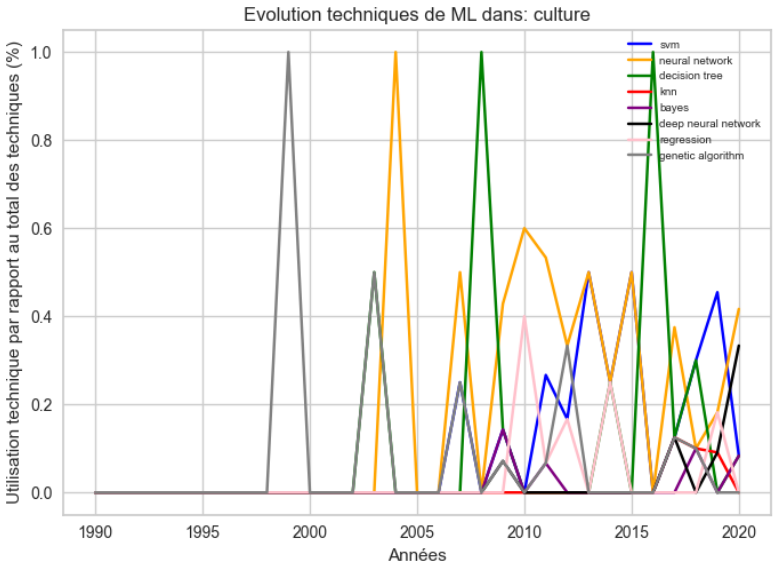


Figure 6-27: Evolution ML dans la culture (Source: Auteurs)

Au vu du trop peu d'articles présents dans le domaine de la ***biologie*** et n'indiquant pas assez les techniques de ML utilisées, il est impossible de donner un sens à l'interprétation du graphe et à l'évolution des techniques (Figure 6-28).

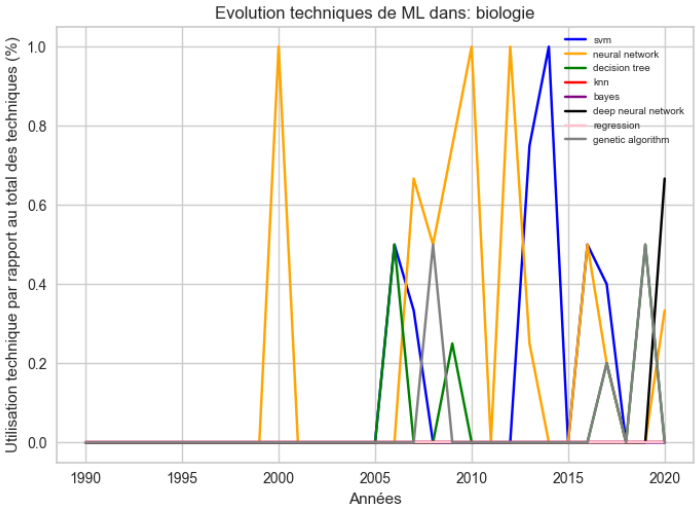


Figure 6-28: Evolution ML dans la biologie (Source: Auteurs)

7 Limites

Une des limites de notre domaine concerne l'analyse manuelle des mots-clés et tokens afin de déterminer si ceux-ci étaient liés à un domaine ou non. Cela est une limite étant donné qu'elle dépend de la connaissance de celui qui a réalisé le travail. C'est pourquoi pour de futures recherches, il serait intéressant d'obtenir des dictionnaires représentant des domaines qui permettraient d'être moins subjectifs dans la méthode de faire ou d'utiliser des articles labellisés où la méthode de classification pourrait être appliquée dessus. Cela permettrait également un gain de temps car l'analyse pourrait se faire de manière automatisée.

Une autre limitation est liée à la disproportion qui existe entre la présence des différents domaines. Des domaines tels que le médical ou la finance sont fortement présents alors qu'il existe des articles liés à l'agriculture qui sont peu présents. Ceux-ci se font absorber dans les domaines de plus grandes importances biaisant certains résultats. Pour remédier à ce problème, il pourrait être intéressant de réaliser une étude qui identifierait les domaines des documents via la classification plutôt que la modélisation de sujet.

Une dernière limite est une limite matérielle. En effet, au vu des capacités de l'ordinateur sur lequel l'analyse s'est faite, il était nécessaire d'utiliser les descriptions plutôt que l'article dans son entièreté pour des questions de performances et de stockage. Afin d'obtenir de meilleurs résultats, il serait intéressant de procéder à une étude sur l'entièreté de l'article mais cela recommande d'avoir une machine plus performante. Cela permettrait d'avoir une meilleure classification des documents dans les sujets car il y aurait beaucoup plus de mots ce qui faciliterait l'identification des mots importants et reflétant le sujet, contrairement à la description où il y a très peu de mots. L'article entier permettrait également de meilleurs résultats sur l'analyse des techniques de ML. En effet, plusieurs auteurs indiquent qu'ils utilisent des techniques de ML dans les mots-clés ou descriptions mais ne mentionnent pas lesquelles. Cela peut donc fausser certains résultats.

8 Conclusion

Pour rappel, notre question de recherche était :

« Comment l'usage des techniques de machine learning a-t-il évolué au sein de différents domaines d'applications ?».

Nous avons pu identifier différents domaines dans la littérature scientifique à l'aide de la technique de modélisation de sujets. Celle-ci nous a permis de faire ressortir 11 sujets interprétables: ***la finance, l'analyse de texte, l'intrusion – détection de défaut, le médical, la manufacture, les maladies cancéreuses et du cœur, l'analyse de mouvement, la gestion d'entreprise, les signaux physiologiques, la biologie et la culture***.

A travers l'évolution des techniques de ML de manière générale et à travers les différents domaines, nous avons pu voir que deux éléments ont eu un impact sur l'usage des techniques. Le premier élément est l'apparition des SVMs vers l'année 2004, qui se révèle être une technique de classification très performante. Les SVMs ont particulièrement été utiles pour la gestion des maladies cancéreuses et cardiaques. Le deuxième élément est l'apparition du deep learning vers 2010 qui influence fortement les différents domaines depuis ces dernières années, particulièrement la détection d'intrusion et de défauts. Au vu des résultats et des performances, nous nous attendons à ce que cette technique soit de plus en plus utilisée.

D'autres éléments externes aux techniques de machine learning ont également eu un impact dans l'usage des techniques de ML. En effet, nous avons pu voir que l'apparition de réseaux sociaux comme Twitter et l'augmentation des éléments sur le web ont cruellement augmenté les publications sur l'analyse de texte et de ce fait l'utilisation du machine learning. De nouvelles technologies comme des capteurs ont également permis d'augmenter la détection d'intrusion et de défauts.

Pour conclure, nous avons pu voir que pour la plupart des domaines, les réseaux de neurones et les SVMs semblaient se distinguer dans l'utilisation. Toutefois, la régression et les arbres de décisions étaient deux méthodes à plus faible utilisation mais constante dans le temps. Cela s'explique par la capacité de ces modèles à pouvoir être interpréter. C'est pour cela qu'ils ont une forte utilisation dans la médecine ou la gestion d'entreprise dont le contact humain est important et dans lequel il y a un besoin d'explications.

9 Annexes

1. Classement Scimago indice H (Source : (SCIMAGO LAB, 2020))

	Title	Type	SJR	↓ H index	Total Docs. (2019)	Total Docs. (3years)	Total Refs. (2019)	Total Cites (3years)	Citable Docs. (3years)	Cites / Doc. (2years)	Ref. / Doc. (2019)
1	IEEE Transactions on Pattern Analysis and Machine Intelligence	journal	7.536 Q1	344	218	626	11845	15167	601	26.69	54.33
2	IEEE Transactions on Neural Networks and Learning Systems	journal	3.555 Q1	196	339	1027	13698	11088	1016	10.52	40.41
3	Pattern Recognition	journal	2.323 Q1	195	381	1142	19783	10074	1115	9.56	51.92
4	Journal of Machine Learning Research	journal	2.219 Q1	188	184	694	8926	3950	687	3.99	48.51
5	Expert Systems with Applications	journal	1.494 Q1	184	725	1877	39322	14829	1860	7.84	54.24

2. Mots reliés Machine learning (Source : (SCIENCE DIRECT, 2021))

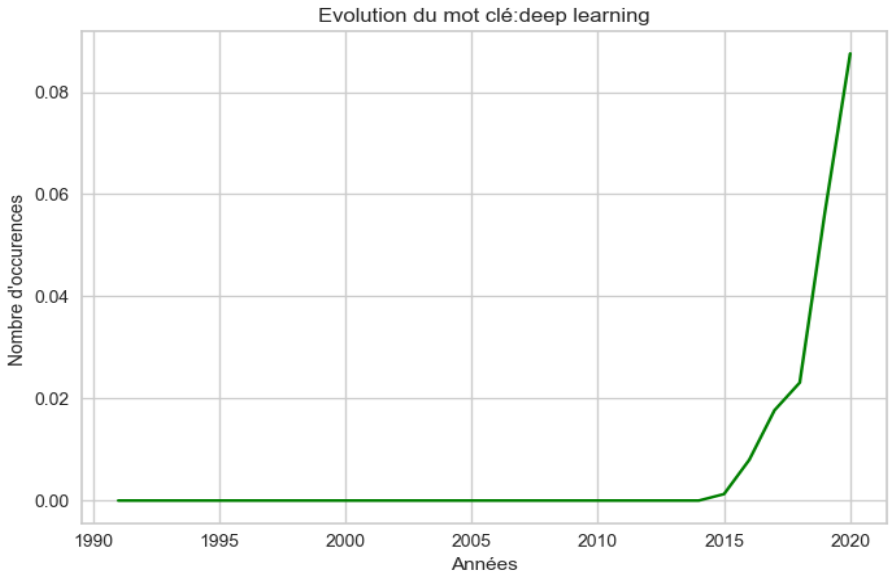
3. Tableau génération mot-clé (Source : Auteurs)

Titre	Description	Mots-clés après génération
Integration of adaptive machine learning and knowledge-based systems for routing and scheduling applications	The combination of good mathematical models, knowledge-based systems, artificial neural networks, and adaptive genetic searches are shown to be synergistic. Practical applications of this combination produces near-optimal results, which none of the individual methods can produce on its own. We have developed XROUTE, a software system that demonstrates an integrated framework for this synergism, in the domain of computer-aided vehicle routing and scheduling problems. The purpose of this system is to assist researchers and decision makers who are applying the mathematical models to a specific routing problem instance by “tuning” the models to the problem description. The neural network modules store knowledge of previously solved problems and their solutions which facilitates the process of arriving at solutions to new problems. The knowledge-based system stores partial solutions from various knowledge sources, like the neural network and genetic algorithm modules, in the working memory and closely supervises the solution process in heuristic mathematical models. XROUTE provides an experimental, exploratory framework that allows many variations, and compares the alternatives on problems with different characteristics. The resultant system is dynamic, expandable, and adaptive and typically outperforms alternative methods in computer-aided vehicle routing.	['mathematical model', 'vehicle routing', 'scheduling problem', 'neural network', 'practical application']

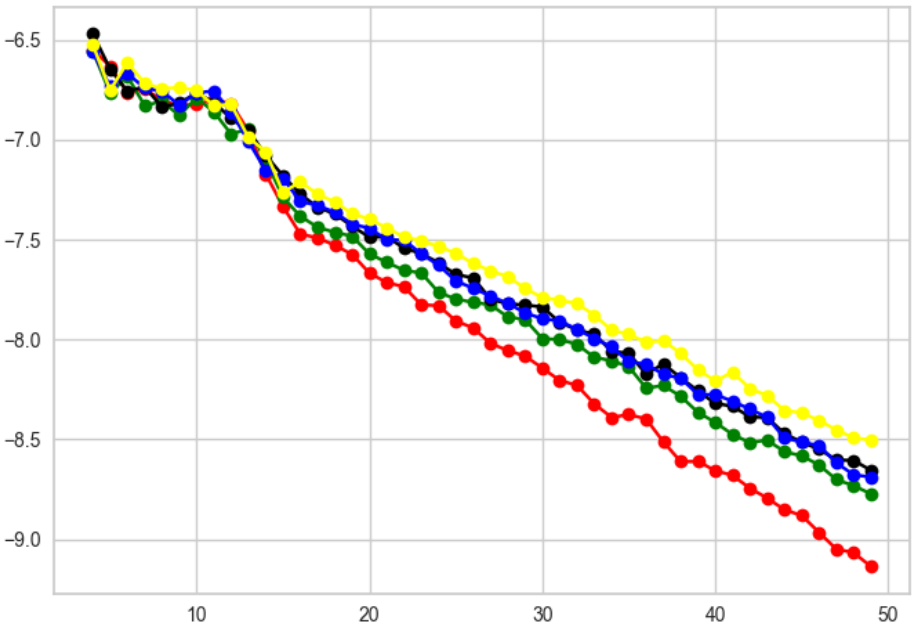
4. Liste des mots-vides (Source : Auteurs)

i,me,my,myself,we,our,ours,ourselves,you,you're,you've,you'll,you'd,your,yours,yourself,yourselves,he,him,his,himself,she,she's,her,hers,herself,it,it's,its,itself,they,them,their,theirs,themselves,what,which,who,whom,this,that,that'll,these,those,am,is,are,was,were,be,been,being,have,has,had,having,do,does,did,doing,a,an,the,and,but,if,or,because,as,until,while,of,at,by,for,with,about,against,between,into,through,during,before,after,above,below,to,from,up,down,in,out,on,off,over,under,again,further,then,once,here,there,when,where,why,how,all,any,both,each,few,more,most,other,some,such,no,nor,not,only,own,same,so,than,too,very,s,t,can,will,just,don,don't,should,should've,now,d,ll,m,o,re,ve,y,ain,aren,aren't,couldn,couldn't,didn,didn't,doesn,doesn't,hadn,hadn't,hasn,hasn't,haven,haven't,isn,isn't,ma,mightn,mightn't,mustn,mustn't,needn,needn't,shan,shan't,shouldn,shouldn't,wasn,wasn't,weren,weren't,won,won't,wouldn,wouldn't,about,again,all,almost,also,although,always,am,among,an,and,another,any,are,as,at,be,because,been,before,being,between,both,but,by,can,could,did,do,does,done,due,during,each,either,enough,especially,etc,ever,for,found,from,further,had,hardly,has,have,having,hence,her,here,him,his,how,however,if,in,into,is,it,its,itself,just,made,mainly,make,might,most,mostly,must,nearly,neither,obtained,of,often,on,onto,or,our,overall,perhaps,quite,rather,really,regarding,said,seem,seen,several,she,should,show,showed,shown,shows,significantly,since,so,some,such,than,that,the,their,theirs,them,then,there,thereby,therefore,these,they,this,those,through,thus,to,too,upon,use,used,using,various,very,was,we,were,what,when,where,whereby,wherein,whether,which,while,whom,whose,why,with,within,without,would,you

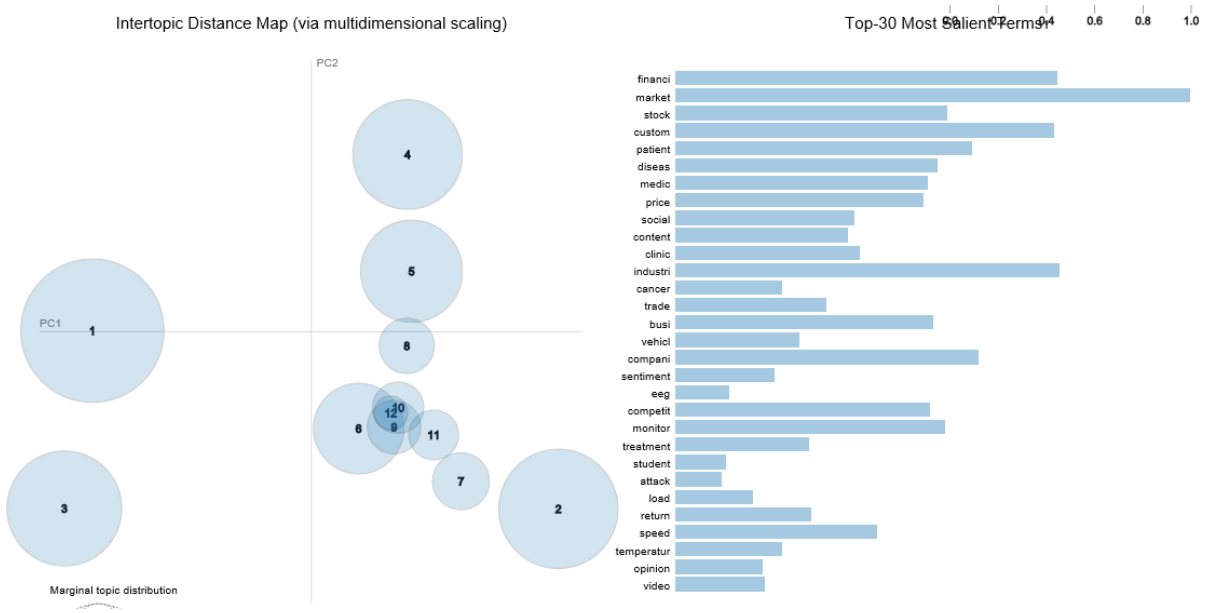
5. Evolution du mot clé *deep learning*. (Source : Auteurs)



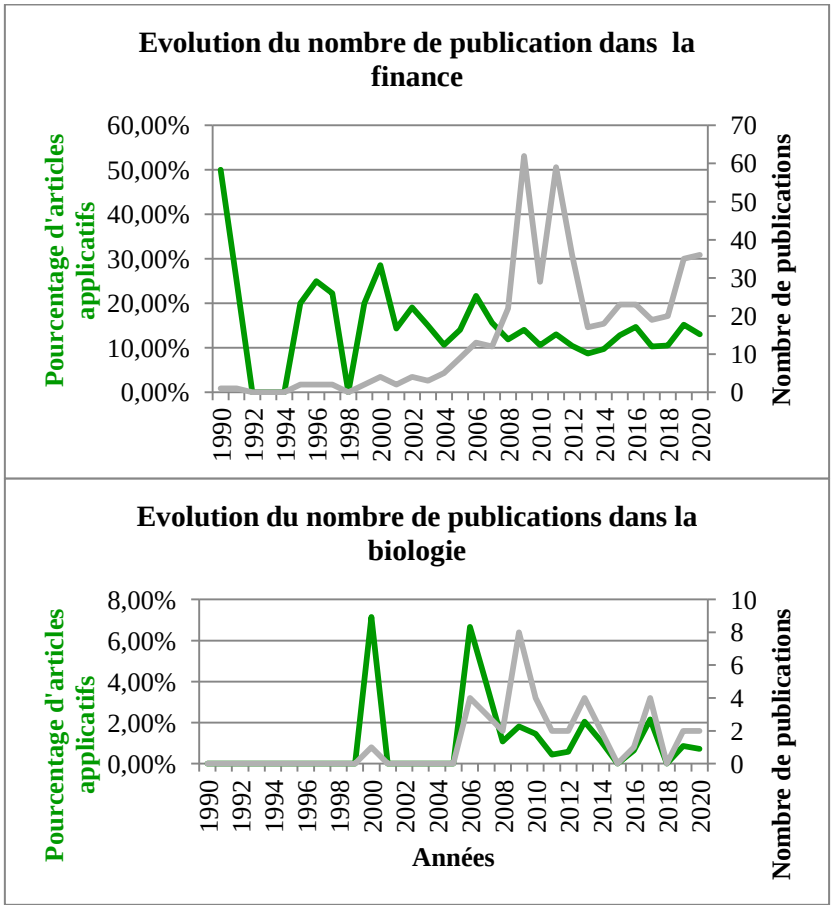
6. Perplexité en fonction du nombre de sujets. (Source : Auteurs)

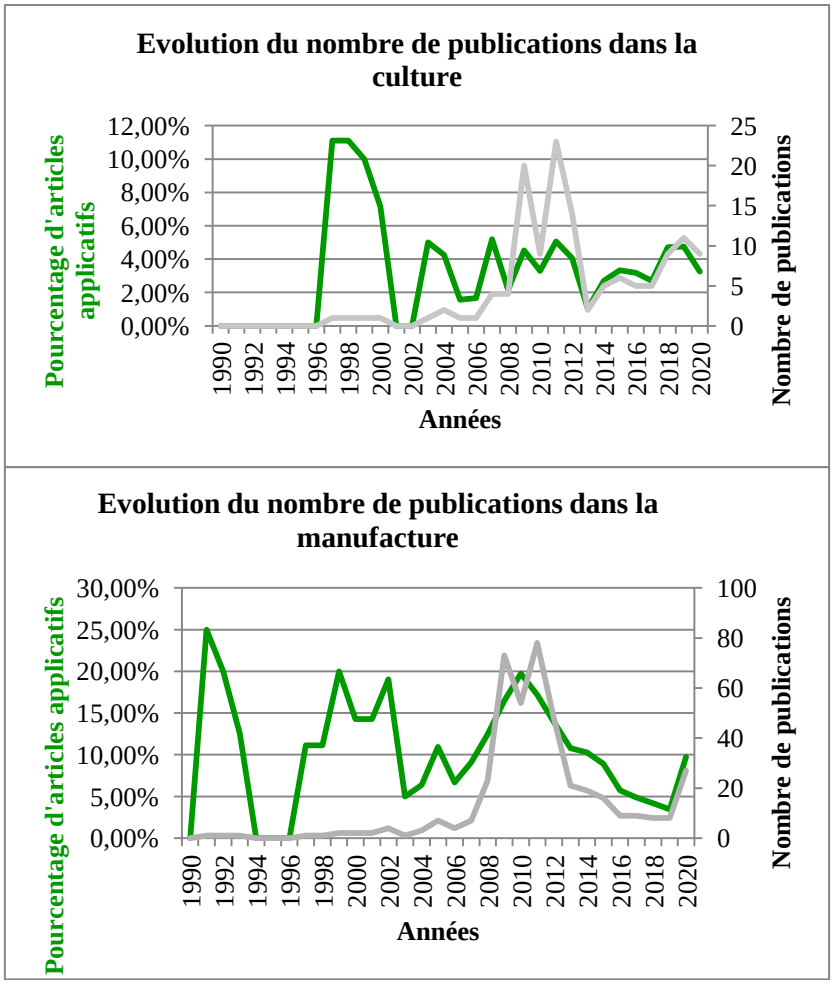


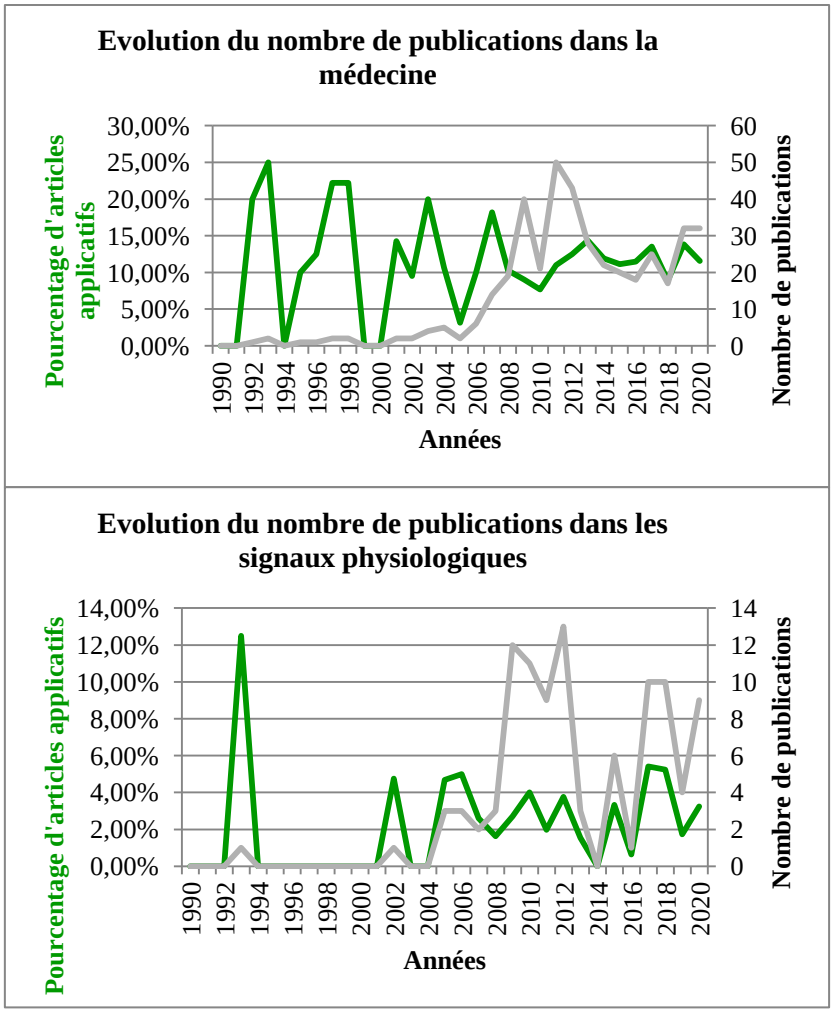
7. Affichage modèle LDA (Source : Auteurs)



8. Evolution du nombre de publications dans différents domaines (Source : Auteurs)




Evolution du nombre de publications dans la culture
Pourcentage d'articles applicatifs
12,00%
10,00%
8,00%
6,00%
4,00%
2,00%
0,00%
25
20
15
10
5
0
Nombre de publications
1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018 2020
Années
Evolution du nombre de publications dans la manufacture
Pourcentage d'articles applicatifs
30,00%
25,00%
20,00%
15,00%
10,00%
5,00%
0,00%
100
80
60
40
20
0
Nombre de publications
1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018 2020
Années


Evolution du nombre de publications dans la médecine
Pourcentage d'articles applicatifs
30,00%
25,00%
20,00%
15,00%
10,00%
5,00%
0,00%
60
50
40
30
20
10
0
Nombre de publications
1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018 2020
Années
Evolution du nombre de publications dans les signaux physiologiques
Pourcentage d'articles applicatifs
14,00%
12,00%
10,00%
8,00%
6,00%
4,00%
2,00%
0,00%
14
12
10
8
6
4
2
0
Nombre de publications
1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018 2020
Années

10 Bibliographie

ASMUSSEN, C., & MOLLER, C. (2019). "Smart literature review: a practical topic modelling approach to exploratory literature review.". *Journal of Big Data, 6*(1), pp. 1-18.

BENGFORT, B., BILBRO, R., & OJEDA, T. (2018). *Applied Text Analysis with Python.* (O. Media, Éd.)

BOSE, I., & MAHAPATRA, R. K. (2001). "Business data mining—a machine learning perspective.". *Information & management, 39*(3), pp. 221-225.

CARBONNEAU, R., LAFRAMBOISE, K., & VAHIDOV, R. (2008). "Application of machine learning techniques for supply chain demand forecasting". *European Journal of Operational Research, 184*(3), pp. 1140-1154.

CENTRE UNIVERSITAIRE DE SANTE DE Mc GILL. (2021). *Quel est votre impact ? En savoir plus sur l'indice h*. Consulté le Mars 18, 2021, sur Bibliothèques du CUSM: https://www.bibliothequescusm.ca/formations-et-conseils/guides-et-tutoriels/quel-est-votre-impact-en-savoir-plus-sur-lindice-h/,

DANG, S., & AHMAD, P. H. (2014). "Text mining: Techniques and its application.". *International Journal of Engineering & Technology Innovations, 1*(4), pp. 866-2348.

DASTILE, X., CELIK, T., & POTSANE, M. (2020). "Statistical and machine learning models in credit scoring: A systematic literature survey.". *Applied Soft Computing, 91*, p. 106263.

DELEN, D., & CROSSLAN, M. D. (2008). "Seeding the survey and analysis of research literature with text mining". *Expert Systems with Applications, 34*(3), pp. 1707-1720.

DEPARTMENT STATISTA RESEARCH. (2019, Juin 6). *Valeur projection marche global intelligence artificielle*. Consulté le Avril 15, 2021, sur Statista: https://fr.statista.com/statistiques/829001/valeur-projection-marche-global-intelligence-artificielle/

ELSEVIER. (2021). *What words are not used in a Scopus search ?* Consulté le Avril 02, 2021, sur Service Elsevier: https://service.elsevier.com/app/answers/detail/a_id/14808/supporthub/scopus/

EUROPEAN CHEMICALS AGENCY (ECHA). (s.d.). *QSAR models*. Consulté le Avril 21, 2021, sur Echa Europea: https://echa.europa.eu/support/registration/how-to-avoid-unnecessary-testing-on-animals/qsar-models

FRENAY, B. (2019). "Introduction and Course overview".

GE, Z., SONG, Z., DING, S. X., & HUANG, B. (2017). "Data mining and analytics in the process industry: The role of machine learning.". *Ieee Access, 5*, pp. 20590-20616.

GITHUB. (2021, Avril 04). *Elsapy*. Récupéré sur github: https://github.com/ElsevierDev/elsapy,

GOMAERE, G. (2019, Mai 20). *Origines-intelligence-artificielle*. Consulté le Avril 15, 2021, sur Journal du CM: https://www.journalducm.com/origines-intelligence-artificielle/

GROUPE MADEINFUTURA. (2020, Novembre 16). *Intelligence artificielle : qu'est-ce que c'est ?* Consulté le Avril 02, 2021, sur Futura-sciences: https://www.futura-sciences.com/tech/definitions/informatique-intelligence-artificielle-555/

GUPTA, V., & LEHAL, G. S. (2009). "A survey of text mining techniques and applications.". *Journal of emerging technologies in web intelligence, 1*, pp. 60-76.

HAO, T., CHEN, X., LI, G., & YAN, J. (2018). "A bibliometric analysis of text mining in medical research.". *Soft Computing, 22*(23), pp. 7875-7892.

KANNAN, S., GURUSAMY, V., VIJAYARANI, S., ILAMATHI, J., & NITHYA, M. (2014). "Preprocessing techniques for text mining.". *International Journal of Computer Science & Communication Networks, 5*(1), pp. 7-16.

KAUSHIk, A., & NAITHANI, S. (2016). "A comprehensive study of text mining approach.". *International Journal of Computer Science and Network Security(IJCSNS), 16*(2), p. 69.

KIM, K., PARK, O., YUN, S., & YUN, H. (2017). "What makes tourists feel negatively about tourism destinations? Application of hybrid text mining methodology to smart destination management.". *Technological Forecasting and Social Change, 123*, pp. 362-369.

KOBAYASHI, V., MOL, S., BERKERS, H., KISMIKÓH, G., & HARTOG, D. (2018). "Text mining in organizational research. Organizational research methods". *D.N, 21*(3), pp. 733-765.

KONONENKO, I. (2001). "Machine learning for medical diagnosis: history, state of the art and perspective". *Artificial Intelligence in Medicine, 23*(1), pp. 89-106.

L, B. (2020, Novembre 18). *Machine Learning : Définition, fonctionnement, utilisations.* Consulté le 04 02, 2021, sur Datascientest: https://datascientest.com/machine-learning-tout-savoir

LIAKOS, K. G., BUSATO, P., MOSHOU, D., PEARSON, S., & BOCHTIS, D. (2018). "Machine learning in agriculture: A review.". *Sensors, 18*(8), p. 2674.

LIN, W. Y., HU, Y. H., & TSAI, C. F. (2011). "Machine learning in financial crisis prediction: a survey". *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42*(4), pp. 421-436.

MITCHELL, T. M. (1997). *"Machine Learning"*. McGraw-Hill.

MORO, S., PIRES, G., RITA, P., & CORTEZ, P. (2019). "A text mining and topic modelling perspective of ethnic marketing research". *Journal of Business Research,, 103*, pp. 275-285.

MOSAVI, A., OZTURK, P., & CHAU, K. W. (2018). "Flood prediction using machine learning models: Literature review.". *Water, 10*(11), p. 1536.

NAUSHAN, H. (2020, Décembre 3). *Topic modeling with latent dirichlet allocation*. Consulté le Mai 08, 2021, sur Towards Datascience: https://towardsdatascience.com/topic-modeling-with-latent-dirichlet-allocation-e7ff75290f8

NGAI, E. W., XIU, L., & CHAU, D. C. (2009). "Application of data mining techniques in customer relationship management: A literature review and classification.". *Expert systems with applications, 36*(2), pp. 2592-2602.

NLTK PROJECT. (2021). *Natural Language Toolkit*. Consulté le Avril 02, 2021, sur NLTK: https://www.nltk.org/

PEJIC-BACH, M., BERTONCEL, T., MESKO, M., & KRSTIC, Z. (2020). "Text mining of industry 4.0 job advertisements". *International journal of information management, 50*, pp. 416-431.

PRINCETON UNIVERSITY. (2010). *About Wordnet*. Consulté le Avril 02, 2021, sur Princeton University: https://wordnet.princeton.edu/

QU, Y., QUAN, P., LEI, M., & SHI, Y. (2019). "Review of bankruptcy prediction using machine learning and deep learning techniques". *Procedia Computer Science, 162*, pp. 895-899.

RESEARCHGATE GMBH. (2008-2021). *profiles*. Consulté le Mars 20, 2021, sur researchgate: https://www.researchgate.net/directory/profiles

SADASTRASOUL, S., GHOLAMIAN, M., SIAMIi, M., & HAJIMOHAMMADI, Z. (2013). "Credit scoring in banks and financial institutions via data mining techniques: A literature review". *Journal of AI and Data Mining, 1*(2), pp. 119-129.

SALLOUM, S. A., AL-EMRAN, M., MONEM, A., & SHAALAN, K. (2017). "A survey of text mining in social media: facebook and twitter perspectives". *Adv. Sci. Technol. Eng. Syst. J, 2*(1), pp. 127-133.

SALLOUM, S., AL-EMRAN, M., MONEM, A., & SHAALAN, K. (2018). "Using text mining techniques for extracting information from research articles". (Springer, Éd.) *Intelligent natural language processing: Trends and Applications*, pp. 373-397.

SCIENCE DIRECT. (2021). *machine learning*. Consulté le Avril 25, 2021, sur sciencedirect: https://www.sciencedirect.com/topics/computer-science/machine-learning

SCIMAGO LAB. (2020, Avril). *Journal Rank*. Consulté le Octobre 8, 2020, sur Scimagojr: https://www.scimagojr.com/journalrank.php?category=1702&order=h&ord=desc

SEVERYN, A., MOSCHITTI, A., URYUPINA, O., PLANK, B., & FILIPPOVA, K. (2016). "Multi-lingual opinion mining on YouTube.". *Information Processing & Management, 52*(1), pp. 46-60.

SHARP, L., AK, R., & HEDBERG JR, T. (2018). "A survey of the advancing use and development of machine learning in smart manufacturing.". *Journal of manufacturing systems, 48*, pp. 170-179.

STREYVERS, M., & GRIFFITHS, T. (2004, Avril). "Finding scientific topics". *Proceedings of the National Academy of Sciences, 1*(2).

SUN, L., & YIN, Y. (2017). "Discovering themes and trends in transportation research using topic modeling". *Transportation Research Part C: Emerging Technologies, 77*, pp. 49-66.

SYAM, N., & SHARMA, A. (2018). "Waiting for a sales renaissance in the fourth industrial revolution: Machine learning and artificial intelligence in sales research and practice.". *Industrial Marketing Management,, 69*, pp. 135-146.

VOYANT, C., NOTTON, G., KALOGIROU, S., NIVET, M. L., PAOLI, C., MOTTE, F., & FOUILLOY, A. (2017). "Machine learning methods for solar radiation forecasting: A review". *Renewable Energy, 105*, pp. 569-582.

WUEST, T., WEIMER, D., IRGENS, C., & THOBEN, K. D. (2016). "Machine learning in manufacturing: advantages, challenges, and applications.". *Production & Manufacturing Research, 4*(1), pp. 23-45.

YANG, H., SPASIC, I., KEANE, J. A., & NENADIC, G. (2009). "A Text Mining Approach to the Prediction of Disease Status from Clinical Discharge Summaries". *Journal of the American Medical Informatics Association, 16*(4), pp. 596-600.

ZHANG, L., TAN, J., HAN, D., & ZHU, H. (2017). "From machine learning to deep learning: progress in machine intelligence for rational drug discovery". *Drug Discovery Today, 22*(11), pp. 1680-1685.