Une revue systématique de la littérature autour du biais, de l'équité et de l'explicabilité

M.L. Ndao[1,2], G. Youness[1,2], N. Niang[2], G. Saporta[2]

[1] Laboratoire LINEACT CESI, Nanterre, IDFC
[2] Laboratoire Cedric-MSDMA, Paris, France

mlndao@cesi.fr ; gyouness@cesi.fr ; ndeye.niang_keita@cnam.fr ; gilbert.saporta@cnam.fr

Résumé

Ce travail fournit une analyse d'une bibliographie autour du biais de l'équité et de l'explicabilité des algorithmes de l'IA entre 2015 et 2022. Par une approche de Traitement Automatique du Langage Naturel, plus précisément la LDA, nous avons extrait 8 sujets traités par cette bibliographie. Une analyse de la popularité de ces sujets a permis de constater une évolution plus rapide du nombre et du pourcentage des publications traitant surtout l'explicabilité et l'équité dans les algorithmes de l'IA. Une comparaison a permis de noter une similarité entre nos résultats et ceux fournis par BERTopic.

Mots-clés

Intelligence Artificielle Explicable (XAI), biais, équité, Traitement Automatique du Langage Naturel (TAL), Latent Dirichlet Allocation (LDA)

Abstract

This work provides an analysis of a bibliography concerning the bias, fairness and explainability of AI algorithms between 2015 and 2022. Using a Natural Language Processing approach, specifically LDA, we extracted 8 topics covered by this bibliography. An analysis of the frequency of these topics showed a faster increase in the number and proportion of publications dealing mainly with explainability and fairness in AI algorithms. A comparison revealed a similarity between our results and those provided by BERTopic.

Keywords

Explainable Artificial Intelligence (XAI), bias, fairness, Natural Language Processing (NLP), Latent Dirichlet Allocation (LDA)

Introduction

Dans un contexte marqué par l'utilisation massive des algorithmes d'apprentissage automatique en Intelligence Artificielle (IA) dans les processus de prise de décision dans presque tous les domaines (finance [31], recommandation [38], santé [26] , etc.) un besoin de confiance en ces algorithmes se pose. Selon Alain Mille et al. (2020) [28], « Nous sommes dans un contexte où les algorithmes de l'IA, initialement destinés à automatiser des tâches mécaniques, s'intéressent à des fonctions cognitives que l'on pensait hors champs de l'automatisation. Ayant eu le statut d'objet de recherche à partir de 1956 (conférence de Dartmouth), l'IA intervient, aujourd'hui, à tous les niveaux de la vie. ». Cependant, de nombreux incidents ont démontré des failles dans ces algorithmes qui sont souvent source de discrimination dans plusieurs domaines comme en reconnaissance faciale, en justice, en recommandation, en recrutement, en banque, en santé, etc. (Google photo [1], COMPAS [2], logiciel de recrutement chez Amazon [3]).

La plupart des algorithmes d'apprentissage automatique (Machine Learning ML) se basent sur des données d'apprentissage susceptibles de contenir un biais : par exemple, une sous représentation d'un groupe d'individus. Ainsi, ce biais pourrait être reconduit dans les prédictions issues de ces algorithmes. Le cas du logiciel de recrutement d'Amazon peut être expliqué par le fait que l'algorithme s'est basé sur les CV collectés depuis plusieurs années et composés essentiellement de CV d'hommes. De surcroît, ces CV ont été sélectionnés par des humains et sont susceptibles d'avoir été choisis de façon biaisée.

Le développement sans précédent des algorithmes de ML dans presque tous les domaines en termes de prise de décisions est conjugué à des failles en termes de biais discriminatoires (non représentativité d'un groupe d'individus comme l'exemple des données du logiciel d'Amazon), d'équité (décision défavorisant un groupe d'individus) et de manque de compréhension des modèles (explicabilité). Cela a provoqué une vague de recommandations de la part de certains organismes tels que la DARPA (Defense Advanced Research Projects Agency) et à l'annonce de l'XAI (eXplainable Artificial Intelligent) en 2016 (D. Gunning et al. 2019) [15].

Depuis cette annonce, on note une forte multiplication des recherches et publications sur l'équité, l'explicabilité et le biais des algorithmes de l'IA. C'est ce qu'on observe en analysant les données de Google Trends sur les tendances de recherches des termes « explainable XAI », « Bias XAI » et « Fairness XAI » (FIGURE 1).

1. https://www.dailymail.co.uk/sciencetech/article
2. ProPublica. 23 mai 2016 ajouter l'article dans ref
3. https://www.assessfirst.com/fr/algorithme-sexiste-amazon/

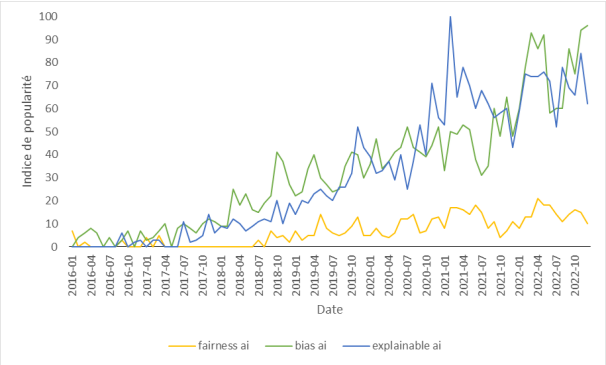

FIGURE 1 – Les tendances de recherches des termes « explainable XAI », « Bias XAI » et « Fairness XAI » dans le monde depuis 2016 selon Google Trends.

Aujourd'hui, une des problématiques autour de la littérature du biais, de l'explicabilité et de l'équité est le nombre important de propositions de modèles d'XAI et de métriques d'équité (plus d'une dizaine de métriques dont certaines sont contradictoires (Mitchell et al., 2021 [30]). Ainsi, un besoin de positionnement des unes par rapport aux autres sur ces différentes propositions se pose. Cela faciliterait l'encadrement de l'utilisation de ces algorithmes afin d'éviter les incidents discriminatoires.

Dans le cadre du biais, nous recensons également un grand nombre de propositions de typologies. Par exemple, Mehrabi et al. (2021)[27] propose une typologie du biais en 3 groupes : des données vers l'algorithme ; de l'algorithme vers les utilisateurs et des utilisateurs vers les données. Tandis que dans Bertail et al. (2019) [2], on retrouve les 3 types de biais suivants : le biais cognitif, le biais statistique et le biais économique. Choisir une typologie du biais parmi les différentes propositions peut être subjectif. Une réorganisation et une recherche de la structure sous-jacente de la bibliographie de l'explicabilité, du biais et de l'équité en IA est nécessaire. C'est l'objectif de ce travail.

Nous proposons une analyse de la structure sous-jacente de la bibliographie autour du biais, de l'équité et de XAI à l'aide d'une approche de Traitement Automatique du Langage Naturel (Natural Language Processing ou NLP), plus précisément le modèle Latent Dirichlet Allocation (LDA, Blei et al., 2003 [3]). D'une part, il s'agit d'identifier les thèmes ou sujets majeurs traités par un ensemble d'articles publiés entre 2015 et 2022. D'autre part, une analyse fine des résultats obtenus aidera à la réorganisation des publications permettant de proposer un état de l'art sur la bibliographie autour du biais de l'équité et de XAI. Enfin, une comparaison de nos résultats et ceux fournis par BERTopic sera effectuée dans le cadre de validation de nos résultats.

La suite du papier est organisée comme suit : la première section est consacrée à une brève présentation des travaux antérieurs qui ont utilisé la LDA pour synthétiser un ensemble d'archives, ainsi qu'à la présentation de l'approche LDA. Ensuite, la section 2 est d'abord dédiée à la présentation de l'ensemble de notre démarche allant de la collecte des données à la modélisation. Par la suite, la deuxième partie de cette section portera sur l'analyse et la discussion des résultats obtenus et leur comparaison avec BERTopic.

1 Topic modeling

1.1 Travaux antérieurs

Le NLP, plus particulièrement le 'topic modeling', est souvent utilisé dans différents domaines selon divers contextes afin de synthétiser, d'organiser ou d'analyser des collections de documents ou d'archives. C'est une approche pertinente dans un contexte de données massives ou big data. Bernadeta et al. 2023 [13] se sont basés sur cette approche pour proposer une analyse synthétique des journaux à propos de la COVID-19 en Suède. Il s'agit d'une description de 6515 articles de journaux publiés entre janvier 2020 et mars 2021. En utilisant l'approche LDA, ils ont pu découvrir les différents sujets traités par ces journaux ainsi que leur évolution dans le temps.

Dans ce même contexte qui est celui la pandémie de la COVID-19, Eren et al. 2020 [9], conscients de la montée rapide du nombre de publications sur la COVID-19, proposent une analyse des archives de la base de données CORD-19 [50]. À ce propos, ils ont utilisé la LDA afin de découvrir la structure en groupes de l'ensemble de ces publications qu'ils visualisent ensuite. Cette étude constitue ainsi une réorganisation des thèmes abordés dans les publications sur la COVID-19 aux USA.

En journalisme, Jacobie at al., 2018 [17] ont également utilisé l'approche LDA pour analyser l'ensemble des publications de The New York Times portant sur la technologie du nucléaire depuis 1945. Cette étude a également prouvé la pertinence de l'approche LDA dans la recherche des sujets sous-jacents à une collection de documents.

Dans le domaine de la maintenance prédictive, Kamal et al. (2021) [33] ont proposé une analyse principalement descriptive de l'ensemble des publications sur l'XAI et la maintenance prédictive entre 2015 et 2021. Il s'agit d'un état de l'art des publications dans ce domaine qui a également permis d'avoir une vue générale, une réorganisation de la bibliographie, mais également une comparaison entre l'explicabilité et la performance des modèles en maintenance prédictive.

Parmi les approches de topic modeling, il y a les approches classiques comme LDA, mais également des approches qui se basent sur des réseaux de neurones comme BERTopic qui est issu de BERT [7] (Bidirectional Encoder Representations from Transformers). BERT est un modèle profond de représentation bidirectionnelle non supervisé du langage développé par Google et qui a donné de bons résultats dans l'extraction de sujets. Grootendorst (2022) [14] s'est basé sur ce modèle pour proposer BERTopic. C'est une extension de BERT en Topic modeling qui se base sur une variation du TF-IDF pour extraire les sujets pertinents. Ce dernier a fourni de bons résultats dans ce domaine [34, 42]. Ainsi, il peut être considéré comme une référence permettant de valider nos résultats. Dans ce travail, nous utiliserons principalement l'approche LDA. BERTopic sera utilisé

dans le cadre de la validation de nos résultats.

1.2 L'approche LDA

Le modèle Latent Dirichlet Allocation (LDA, Blei et al., 2003 [3]) est l'une des techniques de NLP non supervisées les plus connues qui cherchent à découvrir des thématiques ou sujets cachés dans un ensemble de M documents appelé corpus noté D. C'est un modèle probabiliste génératif permettant de trouver la structure sous-jacente d'un ensemble de documents en termes de sujets. La LDA considère le corpus comme un mélange de K sujets décrits chacun par un ensemble de mots auxquels sont associés une probabilité.

L'ensemble des M documents ou encore corpus est représenté par une matrice dite document-mots, souvent creuse, notée $D_{M,N}$ de dimension (M, N) où la cellule (D_i, w_j) correspond à la fréquence du mot w_j dans le document D_i, par exemple :

$$D_{M,N} = \begin{array}{c} \\ D_1 \\ \vdots \\ D_i \\ \vdots \\ D_M \end{array} \begin{array}{c} \begin{array}{cccccc} w_1 & \cdots & w_j & \cdots & w_N \end{array} \\ \begin{bmatrix} 0.3 & & 0 & & 0.2 \\ \cdots & & \cdots & & \\ 0.1 & & 0 & \vdots & 0 \\ & \cdots & & & \vdots \\ 0 & \cdots & 0 & \cdots & 0.01 \end{bmatrix} \end{array}$$

Le nombre de sujets K est choisi *a priori* ou au regard d'un indicateur comme le score de cohérence que l'on définira dans la section suivante.

Partant de $D_{M,N}$, la LDA estime les matrices $\theta_{M,K}$ (documents-sujets) et $\phi_{K,N}$ (sujets-mots) , par une approche itérative.

Dans la matrice $\theta_{M,K}$, $\theta_{m,k}$ correspond à la probabilité que le sujet z_k soit traité dans le document D_m ($\theta_{i\cdot} = \sum_{k=1}^{K} \theta_{ik}$=1).

Ces probabilités sont initialisées par une distribution de Dirichlet de paramètre α ($Dir(\alpha)$). Le résultat est une classification en K clusters où chaque cluster correspond à un sujet. Nous utilisons dans la suite les deux termes sujet ou cluster indifféremment. À partir de $\theta_{M,K}$, on retrouve une partition des documents en K clusters, en affectant chaque document au sujet pour lequel sa probabilité d'appartenance est maximale.

La matrice $\phi_{K,N}$ correspond à la matrice sujets-mots, où ϕ_{kj} correspond à la probabilité que le mot w_j soit dans le sujet z_k. Chaque sujet z_k est décrit par les n mots ayant les plus fortes probabilités ϕ_{kj}, nous les notons $(w_j^k)_{1 \leq j \leq n}$. La matrice $\phi_{K,N}$ est initialisée par une distribution de Dirichlet $Dir(\beta)$. Des exemples de matrices $\theta_{M,K}$ et $\phi_{K,N}$ sont données ci-après :

$$\theta_{M,K} = \begin{array}{c} \\ D_1 \\ \vdots \\ D_i \\ \vdots \\ D_M \end{array} \begin{array}{c} \begin{array}{ccc} z_1 & \cdots & z_K \end{array} \\ \begin{bmatrix} 0 & & 0.2 \\ \vdots & & \vdots \\ 0.1 & & 0.5 \\ & \vdots & \\ 0.6 & \cdots & 0.0 \end{bmatrix} \end{array}$$

$$\phi_{K,N} = \begin{array}{c} \\ z_1 \\ \vdots \\ z_K \end{array} \begin{array}{c} \begin{array}{ccccc} w_1 & \cdots & w_j & \cdots & w_N \end{array} \\ \begin{bmatrix} 0 & & 0 & & 0.3 \\ \cdots & & \cdots & & \\ 0.1 & & 0 & \vdots & 0 \end{bmatrix} \end{array}$$

1.3 Évaluation

Pour évaluer la cohérence quantitative de nos résultats, nous avons utilisé deux métriques appelées scores de cohérence : $UMASS$ (Unnormalized Measures of Association Strength) (Mimno et al., 2011 [29]) et C_V (Coherence Value) score (Röder et al., 2015 [40]). Il s'agit d'indicateurs qui évaluent le degré de similitude sémantique entre les mots les mieux notés dans les sujets en moyenne. Ces deux scores de cohérence sont de bons indicateurs permettant d'évaluer la qualité sémantique des résultats de topic modeling comme LDA ou BERTopic [40]. Le calcul de ces scores de cohérence se base sur la co-occurrence des mots dans l'ensemble des documents et dans chaque sujet et l'information mutuelle.

Par exemple, considérant le sujet z_k et $\epsilon > 0$, le score $UMASS$ est donné par :

$$C_{UMASS} = \frac{2}{N(N-1)} \sum_{i=2}^{N} \sum_{j=1}^{i-1} log \frac{F(w_i^k, w_j^k) + \epsilon}{F(w_j^k)} \quad (1)$$

Ici, $F(w_j^k)$ est le nombre de fois que le mot w_j^k est apparu au moins une fois dans un document et $F(w_i^k, w_j^k)$ le nombre de fois que les w_i^k et w_j^k ont été observés à la fois au sein d'un même document. Le calcul du score C_V se base en plus sur l'information mutuelle ponctuelle normalisée ($NMPI$) donnée par :

$$NMPI(w_i^k, w_j^k) = \frac{log \frac{F(w_i^k, w_j^k) + \epsilon}{F(w_j^k)}}{-log(F(w_i^k, w_j^k + \epsilon))} \quad (2)$$

Plus ces scores sont élevés, meilleure est la qualité des sujets en termes de cohérence. Röder et al., 2015 [40] ont mené une étude comparative des principaux scores de cohérence en les comparant également à l'appréciation humaine. Selon cette étude C_V et $UMASS$ apparaissent comme les meilleures métriques d'évaluation de cohérence.

Dans cette analyse, on utilisera C_V score pour choisir le nombre de sujets. Nous utiliseront les deux métriques pour comparer nos résultats à ceux de BERTopic.

2 Application

Dans cette section, nous commencerons par expliquer notre processus de modélisation depuis la collecte des données. Ensuite, nous présenterons les résultats obtenus à l'issue de cette analyse.

2.1 Processus d'analyse

Dans le cadre de la modélisation, le processus suivant a été suivi (voir FIGURE 2) :

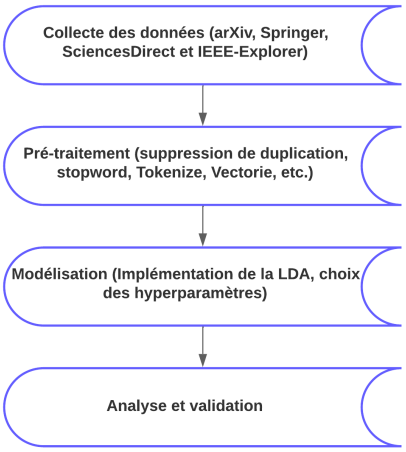

FIGURE 2 – Processus de modélisation.

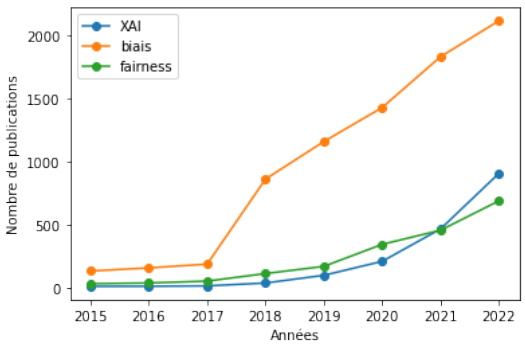

FIGURE 3 – Nombre de publications par thème (XAI, biais, équité) et par année entre 2015 et 2022.

2.1.1 Les données

Cette étude est basée sur les articles publiés sur les quatre plateformes de bases de données suivantes : arXiv, Springer, ScienceDirect et IEEE-Explorer. Sur chaque base de données, nous avons considéré les articles publiés entre 2015 et 2022 avec une recherche séparée sur les méta-données des termes suivants : bias AND (machine learning OR data) ; XAI AND (machine learning OR data) et ; fairness AND (machine learning OR data). Au total, 31 860 articles ont été obtenus. Ensuite, les tâches suivantes ont été réalisées :

- Suppression des duplications : articles ayant les mêmes auteurs, le même titre et le même résumé ;
- Suppression des publications sans résumé ;
- Suppression des articles en d'autres langues que l'anglais.

Par la suite, trois variables binaires ont été créées permettant de vérifier que la publication traite au moins un des trois thèmes : XAI, biais et équité (1 si oui, 0 sinon). Pour chaque thème, les termes suivants ont été considérés :

- Pour XAI : XAI, explainable, explainability, interpretable et interpretability ;
- Pour Biais : bias, harm et disparate ;
- Pour Fairness : fair.

Cette recherche a été faite sur le résumé, le titre et les mots clés de chaque article. Par la suite, seuls les articles ayant traité au moins, un des thèmes a été retenu. Au final, 9 874 publications ont été considérées pour l'étude.

Une analyse du nombre de publications par année montre une augmentation de plus en plus importante de ces dernières, notamment à partir de 2017 (FIGURE 3). Cette situation peut être expliquée par le contexte de cette année (expliquée dans l'introduction) qui a conduit à la création du domaine XAI. Cette analyse montre également que le nombre de publications traitant le biais est plus élevé. En effet, la problématique du biais est présente dans presque tous les domaines. D'où l'importance du nombre de publications qui l'ont traitée.

2.1.2 Pré-traitement

Un pré-traitement a été fait sur les données. Il s'agit de :

- la suppression de mots vides ou 'stopword' qui correspond aux mots qui n'apportent pas d'information tels que : 'the', 'and', etc. Leur suppression accélère l'apprentissage et améliore la précision.
- la tokenisation qui consiste à découper chaque document en une liste de mots appelés tokens. Cette étape conduit à l'obtention d'une matrice documents-termes (matrice d'occurrence).
- la lemmatisation qui consiste à remplacer tous les mots par leur mot-racine. De nombreux mots sont dérivés d'une racine ou d'un mot-racine. (Par exemple, explains, explained $\Rightarrow$ explain) ;
- la normalisation qui consiste à pondérer chaque terme de la matrice d'occurrence. Dans notre cas, nous avons utilisé l'approche tf-idf (Joachims, T. et al., 1996[19]) qui permet d'évaluer l'importance d'un terme dans un document relativement à tous les autres documents.

Ce processus conduit à l'obtention d'une matrice creuse où chaque ligne correspond à un document et chaque colonne correspond à un mot. Suite à ce processus, une analyse du corpus a été faite pour choisir les paramètres optimaux.

2.1.3 Choix du corpus et des paramètres

Pour le choix du corpus, l'analyse est faite sur la concaténation du résumé et des mots-clés. En effet, il n'était pas possible que celle-ci soit faite sur les articles complets car nous disposions à notre niveau uniquement de leurs méta-données : résumé, titre, mots-clés, etc.

Par ailleurs, une analyse séparée a été faite sur le résumé, sur les mots-clés, sur la concaténation du résumé et des mots-clés et sur la concaténation de résumé, mots-clés et titre. L'analyse des résultats obtenus permet de constater que le corpus "concaténation du résumé et les mots-clés" donne de meilleurs résultats qualitatifs (sens des sujets obtenus) et quantitatifs au sens du score de cohérence (voir TABLE 1) qui évalue le degré de similitude sémantique entre les mots les mieux notés dans les sujets en moyenne.

En effet, après pré-traitement, nous avons procédé à un choix des paramètres du modèle K, α et β au regard du

score de cohérence. Ce processus de choix des paramètres est fait en fixant à chaque fois la valeur de K et en faisant varier les valeurs de α et β. Ainsi, nous avons retenu le triplet ($\alpha = 0.91$, $\beta = 0.91$, $K = 8$) (voir TABLE 1) qui donne le meilleur score de cohérence. Dans ce processus de recherche des paramètres optimaux, nous avons également fixé les valeurs de α et β à $0,91$ dans le corpus abstract-keywords puis nous avons varié la valeur de K comme le montre la FIGURE 4. Ainsi, on peut voir que $K = 8$ a le score de cohérence le plus élevé.

Corpus	K	α	β	C_V
Abstract	7	0,31	0,91	0,57
Keywords	9	asymmetric	0,61	0,52
Abstract-keywords	8	0,91	0,91	**0,58**
Abstract-keyw-title	9	0,61	0,91	0,56

TABLE 1 – Résultats de l'analyse des différents corpus.

Notons que l'option 'asymmetric' est une façon d'initialiser des probabilités d'appartenance d'un document à un sujet. Il consiste à initialiser de façon asymétrique ces dernières avec la formule $\frac{1}{topic\_index+\sqrt{(num\_topics)}}$. Pour de raison de comparabilité, ce même nombre de sujets $K = 8$ sera utilisé avec BERTopic.

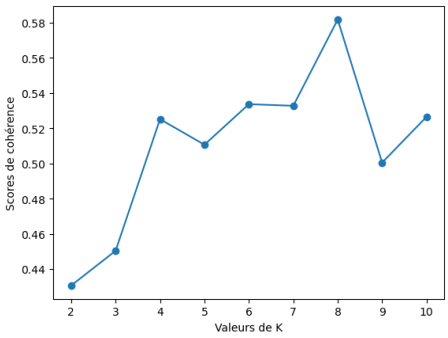

FIGURE 4 – Variation du score de cohérence en fonction du nombre de sujets K lorsque $\alpha = 0.91$ et $\beta = 0.91$ pour le corpus abstract et keywords.

2.2 Résultats et discussions

Dans cette section, nous allons présenter les différents résultats obtenus suite à l'application du modèle LDA dans notre corpus de documents compte tenu des choix de corpus et de hyperparamètres faits plus haut. Dans un premier temps, il s'agira de présenter les 8 sujets obtenus en se basant sur la matrice sujets-termes $\phi_{K,N}$. Dans chaque sujet les mots sont ordonnés suivant l'importance de leurs probabilités. Ensuite, en se basant sur la matrice documents-sujets $\theta_{M,K}$, les documents seront organisés en 8 clusters (sujets).

2.2.1 Analyse des sujets obtenus

Les sujets extraits au moyen de la LDA sont décrits chacun par les 10 mots les plus significatifs en termes de probabilité (TABLE 2). La FIGURE 6 correspond aux pourcentages de documents ayant traité chaque sujet. Par exemple, 13,14% des publications ont traité le sujet1.

L'analyse de ces résultats nous permet de décrire les sujets obtenus de la façon suivante :

- le sujet1, traité par 13,14% des publications, concerne le biais en science cognitive et son impact sur les différents groupes définis par le genre par exemple. La question du biais est très présente dans ce domaine. En effet, les expériences en sciences cognitives sont souvent menées sur des échantillons d'individus. Ainsi, on note de nombreuses publications sur le biais d'échantillonnage qui pourrait limiter la généralisation des résultats issus de ces expériences.
- Le sujet2, traité par le plus faible nombre de publications (2,98%), semble concerner les études de cas autour de l'éthique, la confidentialité des données individuelles (spam, social_medium, véhicule autonome, etc.).
- Le sujet3, traité par 6,77% des publications, porte sur les études de cas en science biologique surtout la génétique et les types de biais rencontrés dans ce domaine.
- Le sujet4, traité par 21,51% des publications, porte sur les données d'images. Il s'agit de d'approches de détection et de classification des images telles que l'apprentissage profond.
- Quant au sujet5, traité par 6,80% des publications, il semble traiter de l'équité, mais dans le cadre du Cloud computing et les objets connectés.
- Quant au sujet6, traité par le plus grand nombre de publications (24,23%), on peut voir qu'il porte sur la confiance en IA à travers l'explicabilité et l'équité des algorithmes de ML. Il s'agit notamment de l'explicabilité des algorithmes dans le cadre de la prise de décisions et des approches (shap, contrefactual) pour garantir l'équité algorithmique.
- Le sujet7, traité par 16,5% des publications, semble porter sur le biais statistique en général. À travers les termes "estimate, forecast, error, parameter" nous pouvons noter qu'il s'agit notamment du biais dans le cadre de l'estimation des paramètres en statistique appliquée.
- Le sujet8, traité par 8,09% des publications, porte sur les études de cas dans le domaine de santé. On note la présence de mots comme "patient, risk, sample, treatment, clinical, etc.".

2.2.2 Analyse des clusters de documents obtenus

L'approche LDA a permis d'organiser les documents en 8 clusters correspondant chacun à un sujet. Cette organisation est faite en se basant sur la matrice de probabilités document-sujet $\theta_{M,K}$. Chaque document est affecté au sujet pour lequel sa probabilité d'appartenance est plus élevée.

Par exemple, le cluster 1 correspond aux publications ayant une plus grande probabilité d'appartenance au sujet1 (FI-

sujet1	sujet2	sujet3	sujet4	sujet5	sujet6	sujet7	sujet8
bias	ethic	cell-coat	feature	fairness	explanation	estimate	patient
cognitive	privacy	gene	image	user	fairness	regression	bias
participant	risk	property	classification	attack	trust	error	risk
attention	policy	structure	detection	cloud_compute	decision	bias	sample
gender	governance	stress	recognition	resource_allocation	human	fault	clinical
stimulus	protection	bias	performance	federate	understand	prediction	vaccine
individual	social_medium	substrate	accuracy	traffic	shap	satellite	climate
group	spam	molecular	semisupervise	agent	counterfactual	forecast	mortality
negative	auto_driving	plasma	task	iot	transparency	performance	disease
social	gdpr	microstructure	cnn	market	interpretable	parameter	treatment

TABLE 2 – Description des sujets par les 10 mots les plus significatifs.

GURE 6).

Visualisation des sujets ou clusters de documents : Pour visualiser les sujets ou clusters de documents dans un espace à deux dimensions, nous avons utilisé l'outil pyLDAviz (Mabey et al., 2021 [25]). Il s'agit d'une approche de visualisation utilisant le positionnement multidimensionnel. L'intérêt de cette visualisation réside dans le fait qu'on arrive à voir :

- la popularité de chaque sujet en termes de nombre de documents l'ayant traité (reflétée par la taille de la surface du cercle) ;
- la similarité entre les différents sujets (reflétée par la proximité entre les cercles).

L'analyse de la distribution des clusters de publications (voir FIGURE 5) montre une bonne séparation de ces premiers (cercles non superposés). Ceci est une garantie de la qualité des résultats obtenus (Mabey et al., 2021 [25]). On peut voir aussi la non-popularité du sujet2 et la proximité entre les sujet7 et sujet4 ainsi que les sujet4 et sujet5.

Nous avons également analysé l'évolution dans le temps du nombre et de la part (en pourcentage) de publications traitant chaque sujet (FIGURE 7 et FIGURE 8). On note que la part de publications qui traitent le sujet6 (en vert) augmente de plus en plus depuis 2018. Cela correspond à une prise de conscience de plus en plus importante sur l'explicabilité des algorithmes de l'IA. Cependant, si la FIGURE 7 semble suivre la même tendance que les données initiales FIGURE 3, la FIGURE 8 a permis d'avoir la part relative de chaque sujet pour chaque année.

On note aussi une montée rapide du nombre de publications ayant traité le sujet4 qui porte sur les modèles de ML en données d'image. En effet, la problématique de l'explicabilité concerne surtout les approches dites "black box" comme les réseaux de neurones. Il s'agit d'approches utilisées surtout dans le cadre des données complexes telles que les données d'image.

Description des clusters de documents : Sur la base de la matrice $\theta_{M,K}$, les documents de chaque cluster ont été ordonnés au regard des probabilités d'appartenance. Chaque

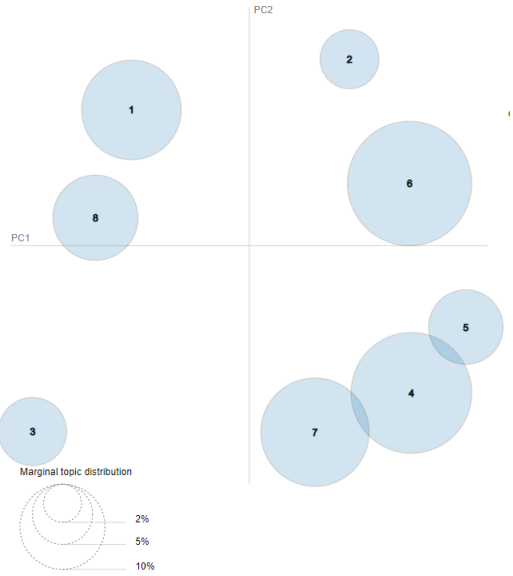

FIGURE 5 – Visualisation de la distribution des clusters par l'outil pyLDAviz (C. Sievert et al, 2014 [46]).

cluster peut être caractérisé par les documents ayant les plus fortes probabilités d'appartenance à ce cluster. Cette description des clusters a permis une organisation de la bibliographie autour du biais, de l'équité et de l'explicabilité en 8 clusters correspondant chacun à un sujet.

En ce qui nous concerne, on s'est surtout intéressé à la description des clusters 4 et 6 parce qu'ils sont plus pertinents par rapport à notre thématique de recherche. De surcroît, une quantification des thèmes recherchés par sujet extrait montre une forte présence de ces deux sujets (voir FIGURE 9).

Cluster 4 : Concernant le cluster 4, les publications les plus significatives sont données dans la TABLE 4 en annexe A. Une analyse approfondie des 21 publications de ce cluster a permis de valider le contenu du sujet correspondant. Il s'agit d'un sujet qui est porté sur l'analyse de données complexes telles que les images et le biais. Parmi les publications les plus significatives de ce cluster, certaines ont

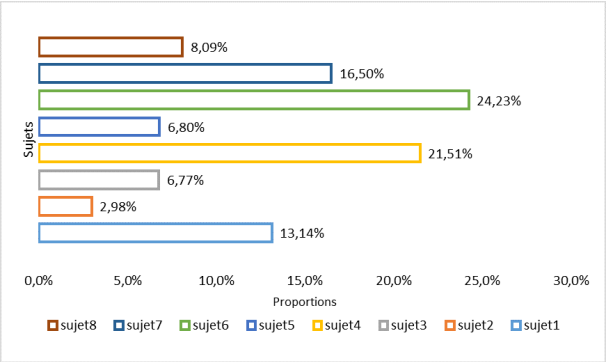

FIGURE 6 – Répartition des publications entre les sujets.

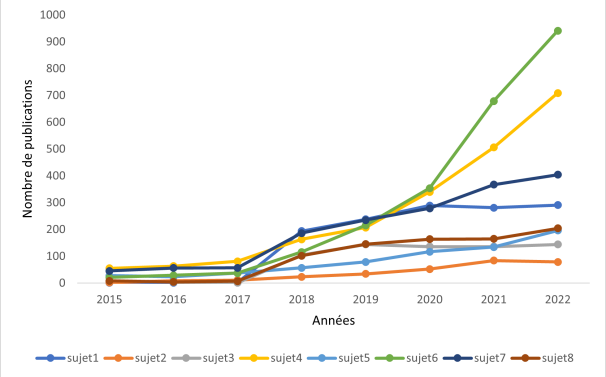

FIGURE 7 – Évolution du nombre de publications par sujet depuis 2015.

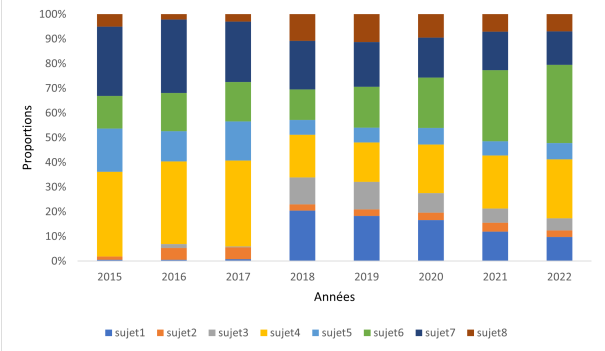

FIGURE 8 – Évolution de la part des publications traitant chaque sujet depuis 2015.

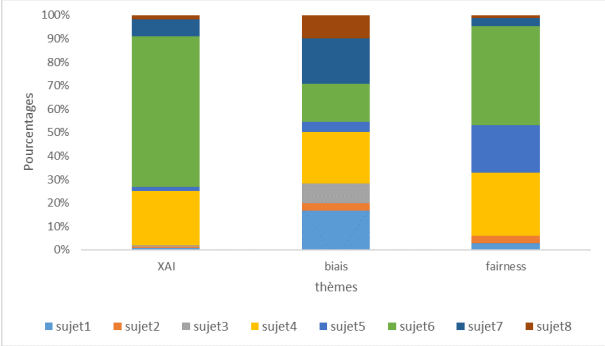

FIGURE 9 – Quantification des thèmes recherchés par sujet. Il s'agit de la répartition des 8 sujets extraits via LDA suivant chaque concept recherché.

analysé le biais selon le type de données (images, graphe, audio, tabulaire). C'est par exemple le cas de Zhengyu Chen et al., 2022 [5]. Constatant que les modèles GNNs (réseaux de neurones en graphes) peuvent être affectés par un potentiel biais lié à une différence de distribution des nœuds dans les données d'entraînement et les données de test, les auteurs ont proposé BA-GNN qui tient compte de cette différence.

Notre analyse a permis également de voir les études sur le biais selon le type de modèle d'analyse utilisé : semi-supervisé (Qiu et al., 2016 [37], Tao et al., 2017 [48])), non supervisé (Dumanvci et al., 2017 [8], Li et al., 2020 [21], Yu et al., 2021[54], Yu et al. [55]) et supervisé (2021[54], Liu et al., 2017 [24]).

Cluster 6 : Une première description a permis de constater que le sujet6 concernait l'équité et l'explicabilité des algorithmes de ML. Lorsqu'on regarde les 11 publications les plus significatives dans le cluster correspondant, on consolide ce constat. Il s'agit d'un sujet traité par des articles qui parlent de l'équité et l'explicabilité (voir TABLE 5, annexe B). Une lecture de ces articles permet de voir une liaison forte de ces deux termes. En effet, ces deux termes sont liés par l'aspect humain, mais également par le fait qu'ils concernent tous les deux directement les décisions prises sur la base des algorithmes de ML. Selon Julie Gerlings et al. (2022) [12], le domaine XAI a été créé dans le but de fournir à l'humain une compréhension des modèles dits de boîtes noires. Cette compréhension conduirait à améliorer la fiabilité des décisions prises sur la base de ces modèles de l'IA.

Si on note, aujourd'hui, un nombre élevé de modèles ou d'approches (appelés modèles XAI) pouvant expliquer les résultats des modèles taxés de "boîtes noires", Jose de Sousa Ribeiro Filho et al., 2022 [39] se demandent si ces explications fournies par l'IA sont conformes aux explications qu'un expert du domaine pourrait fournir. Les résultats de leur analyse ont permis de noter que les explications fournies par les modèles XAI, basées sur les attributs des différentes variables, ne sont pas tout le temps celles fournies par un expert du domaine qui est capable de tenir compte du contexte de l'analyse. Amit Sheth et al., 2021 [45] ont notamment souligné le lien entre la confiance en IA et le niveau d'explicabilité d'un système d'IA. À cet effet, ils soulignent que l'explicabilité ne s'arrête pas aux résultats, mais concerne également d'autres aspects tels que le biais que pourrait contenir les données (par exemple, défaut de représentativité d'un groupe) ou l'équité sur l'acquisition des données.

On note par ailleurs que cette problématique d'explicabilité

concerne tous les types de données et d'approche d'analyse : XAI et Deep Learning (Amit Sheth et al., 2021 [45]), XAI et apprentissage par renforcement (Erika Puiutta et al., 2020 [36]). La limite notée est la non prise en compte de l'aspect humain de façon générale. Cet aspect est essentiel, car pourrait garantir plus d'équité dans les différentes décisions prises sur la base des modèles de l'IA.
Une analyse rapide des autres sujets permet de constater la présence du biais dans de nombreux domaines selon différentes acceptions : (sciences cognitives (sujet1), biologie (sujet3) et santé (sujet8). On a également noté la présence d'un sujet traitant l'éthique et la confidentialité en IA (sujet2), mais également le sujet5 traitant l'équité en Cloud computing.

2.2.3 Comparaison de nos résultats avec ceux de BERTopic

L'application de BERTopic sur nos données montre que nos résultats sont meilleurs que BERTopic au regard de C_V score lorsque HDBSCAN (l'option par défaut) est utilisée comme méthode de clustering(TABLE 3). Cependant, le meilleur C_V score est donné par le modèle BERTopic combiné à un K-moyennes. L'analyse des sujets extraits par ce dernier BERTopic permet de noter une forte similarité entre certains de ces sujets et ceux de notre modèle (TABLE 6 et FIGURE 10) ; on cite : sujet1 BERTopic et sujet4 LDA, sujet2 BERTopic et sujet6 LDA, sujet3 BERTopic et sujet1 LDA).
Cette analyse permet de valider les résultats de notre modèle. Cependant, il faut noter que BERTopic est une approche qualifiée de boîte noire car ayant une structure assez complexe.

	BERT_kmeans	BERT_hdbscan	LDA
UMASS	-0,11	-0,98	-4,31
C_V	0,81	0,57	0,58

TABLE 3 – Scores de cohérence des modèles BERTopic et LDA

Conclusion

Ce travail a permis de classifier la littérature autour du biais, de l'équité et de l'XAI selon 8 sujets : le biais en science cognitive, en santé, en science biologique, en télédétection, l'éthique et la confidentialité des données, les données d'image et les "boîtes noires", l'équité en Cloud computing et enfin l'équité et l'explicabilité des algorithmes de ML. L'analyse de l'évolution des sujets dans le temps permet de noter une évolution plus rapide du nombre et du pourcentage de publications sur les sujets l'équité et l'explicabilité des algorithmes de l'IA.
L'extraction des sujets traités de la bibliographie sur le biais,l'équité et l'XAI a permis de réorganiser de cette bibliographie. Nous avons constaté que cette bibliographie couvre un grand nombre de domaines notamment en ce qui concerne le biais. Il a permis aussi d'extraire un sujet portant exclusivement sur les données complexes comme les images, ainsi que les modèles adaptés à leur analyse comme les CNN qui ont fortement impacté le domaine de l'IA. Ainsi, l'utilisation des approches de traitement de langage naturel pour synthétiser, résumer, et même organiser une bibliographie reste très utile dans un contexte des données massives (big data) où un besoin d'analyse systématique se pose de plus en plus. En effet, cela peut être utile pour organiser une bibliographie en permettant d'aborder de manière directe les principaux sujets d'intérêt. En ce qui nous concerne, la classification obtenue permettra d'organiser notre bibliographie pour une meilleure exploitation de celle-ci.
Cependant, il est important de noter que cette analyse se base sur un échantillon de l'ensemble des publications autour du biais de l'équité et de l'explicabilité. En effet, en dehors des 4 plateformes choisies, il en existe d'autres telles que IJCAI et ECAI. Une analyse d'une plus large base de données pourrait faire ressortir d'autres sujets aussi importants que ceux obtenus. Dans cette analyse, nous avons aussi fait une comparaison rapide entre nos résultats et ceux de BERTopic. Puisque BERTopic fournit de bons résultats dans ce domaine, dans nos travaux futurs, nous souhaitons analyser davantage nos documents à l'aide de ce modèle afin d'améliorer nos résultats.

Références

[1] Kiana Alikhademi, Emma Drobina, Diandra Prioleau, Brianna Richardson, Duncan Purves, and Juan E Gilbert. A review of predictive policing from the perspective of fairness. *Artificial Intelligence and Law*, pages 1–17, 2022.

[2] Patrice Bertail, David Bounie, Stéphan Clémençon, and Patrick Waelbroeck. Algorithmes : biais, discrimination et équité. *NA*, 2019.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan) :993–1022, 2003.

[4] Petra Budikova, Michal Batko, and Pavel Zezula. Conceptrank for search-based image annotation. *Multimedia Tools and Applications*, 77 :8847–8882, 2018.

[5] Zhengyu Chen, Teng Xiao, and Kun Kuang. Ba-gnn : On learning bias-aware graph neural network. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pages 3012–3024. IEEE, 2022.

[6] Ziheng Chen and Jiangtao Ren. Multi-label text classification with latent word-wise label information. *Applied Intelligence*, 51 :966–979, 2021.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*, 2018.

[8] Sebastijan Dumančić and Hendrik Blockeel. An expressive dissimilarity measure for relational clustering using neighbourhood trees. *Machine learning*, 106 :1523–1545, 2017.

[9] Maksim Ekin Eren, Nick Solovyev, Edward Raff, Charles Nicholas, and Ben Johnson. Covid-19 kaggle literature organization. In *Proceedings of the ACM Symposium on Document Engineering 2020*, pages 1–4, 2020.

[10] Zhenyong Fu, Tao Xiang, Elyor Kodirov, and Shaogang Gong. Zero-shot learning on semantic class prototype graph. *IEEE transactions on pattern analysis and machine intelligence*, 40(8) :2009–2022, 2017.

[11] Sainyam Galhotra, Romila Pradhan, and Babak Salimi. Explaining black-box algorithms using probabilistic contrastive counterfactuals. In *Proceedings of the 2021 International Conference on Management of Data*, pages 577–590, 2021.

[12] Julie Gerlings, Millie Søndergaard Jensen, and Arisa Shollo. Explainable ai, but explainable to whom ? an exploratory case study of xai in healthcare. *Handbook of Artificial Intelligence in Healthcare : Vol 2 : Practicalities and Prospects*, pages 169–198, 2022.

[13] Bernadeta Griciūtė, Lifeng Han, Alexander Koller, and Goran Nenadic. Topic modelling of swedish newspaper articles about coronavirus : a case study using latent dirichlet allocation method. *arXiv preprint arXiv :2301.03029*, 2023.

[14] Maarten Grootendorst. Bertopic : Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv :2203.05794*, 2022.

[15] David Gunning and David Aha. Darpa's explainable artificial intelligence (xai) program. *AI magazine*, 40(2) :44–58, 2019.

[16] Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Autonovel : Automatically discovering and learning novel visual categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10) :6767–6781, 2021.

[17] Carina Jacobi, Wouter Van Atteveldt, and Kasper Welbers. Quantitative analysis of large amounts of journalistic texts using topic modelling. In *Rethinking Research Methods in an Age of Digital Journalism*, pages 89–106. Routledge, 2018.

[18] Weina Jin, Jianyu Fan, Diane Gromala, Philippe Pasquier, and Ghassan Hamarneh. Euca : The end-user-centered explainable ai framework. *arXiv preprint arXiv :2102.02437*, 2021.

[19] Thorsten Joachims. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. Technical report, Carnegie-mellon univ pittsburgh pa dept of computer science, 1996.

[20] Michail Kaseris, Ioannis Mademlis, and Ioannis Pitas. Adversarial unsupervised video summarization augmented with dictionary loss. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2683–2687. IEEE, 2021.

[21] Peizhao Li, Han Zhao, and Hongfu Liu. Deep fair clustering for visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9070–9079, 2020.

[22] Ruihui Li, Jianrui Cai, Hanling Zhang, and Taihong Wang. Aggregating complementary boundary contrast with smoothing for salient region detection. *The Visual Computer*, 33 :1155–1167, 2017.

[23] Tao Lian, Lin Du, Mingfu Zhao, Chaoran Cui, Zhumin Chen, and Jun Ma. Evaluating and improving the interpretability of item embeddings using item-tag relevance information. *Frontiers of Computer Science*, 14 :1–16, 2020.

[24] Meng Liu, Chang Xu, Yong Luo, Chao Xu, Yonggang Wen, and Dacheng Tao. Cost-sensitive feature selection by optimizing f-measures. *IEEE Transactions on Image Processing*, 27(3) :1323–1335, 2017.

[25] Ben Mabey. pyldavis documentation, 2021.

[26] Arjun K Manrai, Birgit H Funke, Heidi L Rehm, Morten S Olesen, Bradley A Maron, Peter Szolovits, David M Margulies, Joseph Loscalzo, and Isaac S Kohane. Genetic misdiagnoses and the potential for health disparities. *New England Journal of Medicine*, 375(7) :655–665, 2016.

[27] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6) :1–35, 2021.

[28] Alain Mille, Rémy Chaput, and Amélie Cordier. *Une perspective historique sur l'IA explicable Document préparatoire à un tutorial AFIA juillet 2020*. PhD thesis, LIRIS UMR 5205 CNRS/INSA de Lyon/Université Claude Bernard Lyon 1/Université …, 2020.

[29] David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 262–272, 2011.

[30] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Algorithmic fairness : Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8 :141–163, 2021.

[31] Amitabha Mukerjee, Rita Biswas, Kalyanmoy Deb, and Amrit P Mathur. Multi–objective evolutionary algorithms for the risk–return trade–off in bank loan management. *International Transactions in operational research*, 9(5) :583–597, 2002.

[32] Oliver Nina, Jamison Moody, and Clarissa Milligan. A decoder-free approach for unsupervised clustering and manifold learning with random triplet mining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[33] Ahmad Kamal Bin Mohd Nor, Srinivasa Rao Pedapait, and Masdi Muhammad. Explainable ai (xai) for phm of industrial asset : A state-of-the-art, prisma-compliant systematic review. *arXiv preprint arXiv :2107.03869*, 2021.

[34] Bayode Ogunleye, Tonderai Maswera, Laurence Hirsch, Jotham Gaudoin, and Teresa Brunsdon. Comparison of topic modelling approaches in the banking context. *Applied Sciences*, 13(2), 2023.

[35] Alvin Poernomo and Dae-Ki Kang. Biased dropout and crossmap dropout : learning towards effective dropout regularization in convolutional neural network. *Neural networks*, 104 :60–67, 2018.

[36] Erika Puiutta and Eric MSP Veith. Explainable reinforcement learning : A survey. In *Machine Learning and Knowledge Extraction : 4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2020, Dublin, Ireland, August 25–28, 2020, Proceedings 4*, pages 77–95. Springer, 2020.

[37] Zhicong Qiu, David J Miller, and George Kesidis. A maximum entropy framework for semisupervised and active learning with unknown and label-scarce classes. *IEEE transactions on neural networks and learning systems*, 28(4) :917–933, 2016.

[38] Inioluwa Deborah Raji and Joy Buolamwini. Actionable auditing : Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 429–435, 2019.

[39] José Ribeiro, Níkolas Carneiro, and Ronnie Alves. Black box model explanations and the human interpretability expectations–an analysis in the context of homicide prediction. *arXiv preprint arXiv :2210.10849*, 2022.

[40] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408, 2015.

[41] Maria Sahakyan, Zeyar Aung, and Talal Rahwan. Explainable artificial intelligence for tabular data : A survey. *IEEE Access*, 9 :135392–135422, 2021.

[42] Vasudeva Raju Sangaraju, Bharath Kumar Bolla, Deepak Kumar Nayak, and Jyothsna Kh. Topic modelling on consumer financial protection bureau data : An approach using bert based embeddings. *arXiv preprint arXiv :2205.07259*, 2022.

[43] Teresa Scantamburlo. Non-empirical problems in fair machine learning. *Ethics and Information Technology*, 23(4) :703–712, 2021.

[44] Gesina Schwalbe and Bettina Finzel. A comprehensive taxonomy for explainable artificial intelligence : a systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*, pages 1–59, 2023.

[45] Amit Sheth, Manas Gaur, Kaushik Roy, and Keyur Faldu. Knowledge-intensive language understanding for explainable ai. *IEEE Internet Computing*, 25(5) :19–24, 2021.

[46] Carson Sievert and Kenneth Shirley. Ldavis : A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, pages 63–70, 2014.

[47] S Regina Lourdhu Suganthi, M Hanumanthappa, and S Kavitha. Event image classification using deep learning. In *2018 International Conference on Soft-computing and Network Security (ICSNS)*, pages 1–8. IEEE, 2018.

[48] Hong Tao, Chenping Hou, Feiping Nie, Jubo Zhu, and Dongyun Yi. Scalable multi-view semi-supervised classification via adaptive regression. *IEEE Transactions on Image Processing*, 26(9) :4283–4296, 2017.

[49] Chen Wang, Chengyuan Deng, and Vladimir Ivanov. Sag-vae : End-to-end joint inference of data representations and feature relations. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2020.

[50] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, et al. Cord-19 : The covid-19 open research dataset. *ArXiv*, 2020.

[51] Yali Wang, Lei Zhou, and Yu Qiao. Temporal hallucinating for action recognition with few still images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5314–5322, 2018.

[52] Yulong Wang, Wei Yang, and Haoxin Zhang. Deep learning single logo recognition with data enhancement by shape context. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2018.

[53] Yao Xie, Ge Gao, and Xiang'Anthony' Chen. Outlining the design space of explainable intelligent systems for medical diagnosis. *arXiv preprint arXiv :1902.06019*, 2019.

[54] Heng Yu, Haoran Luo, Yuqi Yi, and Fan Cheng. A2r2 : robust unsupervised neural machine translation with adversarial attack and regularization on representations. *IEEE Access*, 9 :19990–19998, 2021.

[55] Lingli Yu, Xumei Xia, and Kaijun Zhou. Traffic sign detection based on visual co-saliency in complex scenes. *Applied Intelligence*, 49 :764–790, 2019.

[56] Kai Zhao, Qi Han, Chang-Bin Zhang, Jun Xu, and Ming-Ming Cheng. Deep hough transform for semantic line detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9) :4793–4806, 2021.

A Publications ayant une probabilité d'appartenance supérieure à 0.99 au sujet 4

Références	Titres (en anglais)	Probabilités
[37]	A Maximum Entropy Framework for Semisupervised and Active Learning With Unknown and Label-Scarce Classes	0,9944
[35]	Biased Dropout and Crossmap Dropout : Learning towards effective Dropout regularization in convolutional neural network	0,9940
[56]	Deep Hough Transform for Semantic Line Detection	0,9940
[51]	Temporal Hallucinating for Action Recognition with Few Still Images	0,9938
[23]	Evaluating and improving the interpretability of item embeddings using item-tag relevance information	0,9934
[8]	An expressive dissimilarity measure for relational clustering using neighbourhood trees	0,9929
[10]	Zero-Shot Learning on Semantic Class Prototype Graph	0,9925
[22]	Aggregating complementary boundary contrast with smoothing for salient region detection	0,9921
[49]	SAG-VAE : End-to-end Joint Inference of Data Representations and Feature Relations	0,9919
[52]	Deep Learning Single Logo Recognition with Data Enhancement by Shape Context	0,9919
[24]	Cost-Sensitive Feature Selection by Optimizing F-Measures	0,9914
[47]	Event Image Classification using Deep Learning	0,9913
[48]	Scalable Multi-View Semi-Supervised Classification via Adaptive Regression	0,9913
[20]	Adversarial Unsupervised Video Summarization Augmented With Dictionary Loss	0,9912
[16]	AutoNovel : Automatically Discovering and Learning Novel Visual Categories	0,9912
[55]	Traffic sign detection based on visual co-saliency in complex scenes	0,9911
[32]	A Decoder-Free Approach for Unsupervised Clustering and Manifold Learning with Random Triplet Mining	0,9910
[4]	ConceptRank for search-based image annotation	0,9910
[6]	Multi-label text classification with latent word-wise label information	0,9909
[54]	Cross-View Asymmetric Metric Learning for Unsupervised Person Re-Identification	0,9905
[5]	BA-GNN : On Learning Bias-Aware Graph Neural Network	0,9904

TABLE 4 – Les publications ayant une probabilité d'appartenance supérieure à 0.99 au sujet 4

B Publications ayant une probabilité d'appartenance supérieure à 0.99 au sujet 6

Références	Titres (en anglais)	Probabilités
[39]	Black Box Model Explanations and the Human Interpretability in the Context of Homicide Prediction	0,9922
[43]	Non-empirical problems in fair machine learning	0,9922
[45]	Knowledge-Intensive Language Understanding for Explainable AI	0,9912
[11]	Explaining Black-Box Algorithms Using Probabilistic Contrastive Counterfactuals	0,9910
[1]	A review of predictive policing from the perspective of fairness	0,9910
[18]	EUCA : the End-User-Centered Explainable AI Framework	0,9908
[41]	Explainable Artificial Intelligence for Tabular Data : A Survey	0,9908
[12]	Explainable AI, but explainable to whom ?	0,9906
[36]	Explainable Reinforcement Learning : A Survey	0,9902
[53]	Outlining the Design Space of Explainable Intelligent Systems for Medical Diagnosis	0,9900
[44]	A Comprehensive Taxonomy for Explainable Artificial Intelligence : A Systematic Survey of Surveys on Methods and Concepts	0,9900

TABLE 5 – Les publications ayant une probabilité d'appartenance supérieure à 0.99 au sujet 6

C Description des sujets fournis par BERTopic

sujet1	sujet2	sujet3	sujet4	sujet5	sujet6	sujet7	sujet8
algorithm	bias	bias	bias	bias	bias	algorithm	receptor
model	fairness	attentional	patient	exchange	estimation	resource	cell
feature	algorithm	participant	disease	magnetic	model	traffic	signaling
neural	system	cognitive	risk	exchange bias	error	proposed	gene
method	user	attention	clinical	field	correction	detection	protein
proposed	model	attentional bias	result	film	estimate	fairness	agonist
image	decision	stimulus	study	substrate	result	network	biased
classification	social	result	medical	property	measurement	performance	codon
performance	research	task	model	coating	proposed	attack	ligand
problem	information	negative	method	device	estimator	allocation	pathway

TABLE 6 – Description des 8 sujets fournis par BERTopic_kmeans

D Correspondance entre les 8 sujets LDA et les 8 sujets BERT.

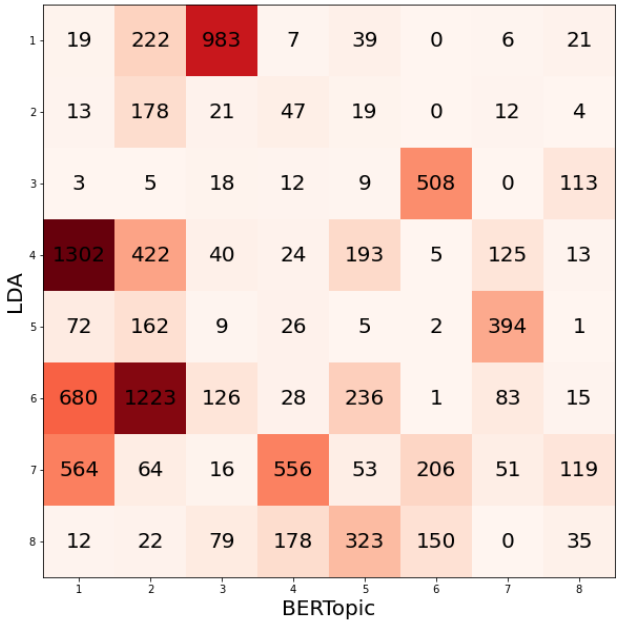

FIGURE 10 – Matrice de confusion entre les clusters de documents des modèles LDA et BERTopic. Cette matrice montre la correspondance quantitative entre les clusters de documents obtenus par les deux approches : LDA et BERTopic.