Consensus de partitions en NLP pour une revue systématique de la littérature autour de l'XAI du biais et de l'équité

Mouhamadou-Lamine Ndao, Ndèye Niang, Genane Youness, Gilbert Saporta

Consensus de partitions en NLP pour une revue systématique de la littérature autour de l'XAI du biais et de l'équité

M.L. Ndao[1,2], N. Niang[2], G. Youness[1,2], 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 présente une analyse comparative d'une bibliographie autour du biais de l'équité et de l'explicabilité des algorithmes de l'IA entre 2015 et 2022. Par trois approches de Traitement Automatique du Langage Naturel (LDA, NMF et k-SVD), nous avons extrait différents sujets traités par cette bibliographie. Ces trois approches nous ont également fourni trois partitions. Dans l'optique d'éviter de faire un choix entre ces partitions, nous avons proposé une synthèse de ces trois partitions par une approche de consensus pondérée.

Mots-clés

Intelligence Artificielle eXplicable (XAI), Traitement Automatique du Langage Naturel (TAL), Non-negative matrix factorization (NMF), Consensus de partitions pondéré

Abstract

This work provides a comparative analysis of a bibliography around fairness bias and explainability of AI algorithms between 2015 and 2022. Through three approaches of Natural Language Processing (LDA, NMF et k-SVD), we extracted different topics covered by this bibliography. These three approaches also provided us with three partitions. In order to avoid making a choice between these partitions, we proposed a synthesis of these three partitions by a weighted consensus approach.

Keywords

eXplainable Artificial Intelligence (XAI), Natural Language Processing (NLP), Non-negative matrix factorization (NMF), Weighted consensus

Introduction

La problématique de l'équité, du biais et de l'équité en apprentissage automatique (Machine Learning ML) est de plus en plus présent. Ceci est lié à de nombreuses 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]). Étant donné que la plupart des algorithmes d'apprentissage automatique (Machine Learning ML) établissent des règles sur la base des données d'apprentissage susceptibles de présenter un biais, il en est de même des prédictions issues de ces algorithmes. Ce contexte a provoqué une vague de recommandations de la part de certains organismes tels que la DARPA (Defense Advanced Research Projects Agency). On assiste à cet effet à l'annonce du concept d'XAI (eXplainable Artificial Intelligent) en 2016 (D. Gunning et al. 2019) [5]. Ce concept met en avant la compréhension par l'humain des décisions prises sur la base des algorithmes de l'IA.
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).

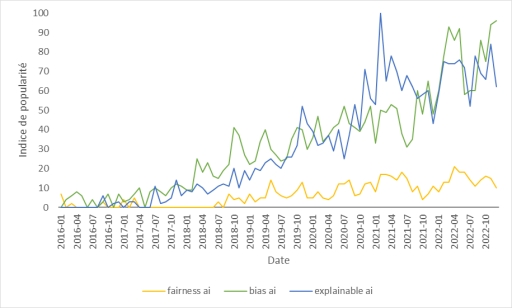

FIGURE 1 – Les tendances de recherches des termes « explainable XAI », « Bias XAI » et « Fairness XAI » dans le monde depuis 2016 selon Google Trends.

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 de 400 références citées par Barredo Arrieta et al., 2020 [1]) dont certaines sont contradictoires (Mitchell et al., 2021 [11]). Ainsi, une réorganisation et une recherche de la structure sous-jacente de la bibliographie de

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/

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 de approche de Traitement Automatique du Langage Naturel (Natural Language Processing ou NLP) non supervisée. Notre approche consistera a utilisé 3 modèles d'analyse : Latent Dirichlet Allocation (LDA, Blei et al., 2003 [2]); NMF et k-SVD. Ensuite, par une approche de consensus de partitions pondérés WNMF, on regardera le compromis entre ces trois modèles d'analyse.

Le reste du papier est organisé comme suit : la première section est consacrée à une brève présentation d'une part des travaux antérieurs qui ont utilisé une approche NLP pour synthétiser un ensemble d'archives et d'autre part les travaux sur le consensus de partitions. Ensuite, la section 2 est dédiée à la présentation de l'ensemble de notre démarche allant de la collecte des données à la modélisation. La deuxième partie de cette section portera sur l'analyse et la discussion des résultats obtenus.

1 Méthodologie

1.1 Topic Modeling

Le Topic modeling est une approche d'apprentissage automatique non supervisée qui est souvent utilisée 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. En effet, elle permet de retrouver une structure sous-jacente d'une collection de documents (partition) en extrayant les sujets liés à chacun des sous-ensembles de cette structure. Il existe de nombreuses approches de topic modeling. Dans le cadre de ce travail, nous nous sommes particulièrement intéressé à trois modèles : le modèle Latent Dirichlet Allocation (LDA, Blei et al., 2003 [2]); la SVD tronqué (appelé également k-SVD) [6] et la Non-negative matrix factorization [9]. Ces trois approches permettent d'avoir : d'une part une relation sujets-mots et d'autre part une relation documents sujets qui conduit à une partition des documents. Dans cette section on se limite à la présentation de l'approche LDA qui est notre modèle de référence.

1.1.1 Principe de LDA

Le modèle Latent Dirichlet Allocation (LDA, Blei et al., 2003 [2]) est 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. Il 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 sparse, 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 :

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

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 $D_{M,N}$, toutes les trois approches estiment les matrices $\theta_{M,K}$ (documents-sujets) et $\phi_{K,N}$ (sujets-mots).

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 ($\theta_{i.} = \sum_{k=1}^{K} \theta_{ik}$=1).

Le résultat est une classification floue 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 :

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

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

Pour évaluer la qualité des sujets obtenues plusieurs mesures sont classiquement utilisées. Il s'agit des indices : Umass (Université du Massachusetts) [2], CV (Coherence value) [10], UCI (Ultra-Compactness Index) [10] et NPMI (Normalized Pointwise Mutual Information) [3]. Parmi ces métriques, nous allons utiliser le score de cohérence CV pour choisir le nombre de sujets K optimal.

En plus de l'approche LDA, nous utiliserons deux autres méthodes : NMF et k-SVD obtenant ainsi 3 partitions éventuellement de qualité différentes. Dans ce travail, nous proposons de synthétiser ces trois à travers une approche de type ensemble afin d'avoir une seule partition des documents.

1.2 Consensus de partitions

Le problème de la combinaison de plusieurs partitions d'un ensemble d'objets (ou d'individus) en une seule partition,

connu également sous le nom de consensus de partitions ou agrégation de partitions, consiste à identifier une partition compromis d'un ensemble de partitions obtenues sur le même ensemble d'observations [4, 12].

Le principe des méthodes de consensus est de trouver la partition compromis des partitions séparées appelées partitions contributives. Cette dernière partition doit être la plus similaire aux partitions contributives. Plusieurs méthodes ont été proposées. Elles peuvent être regroupées en trois grandes familles : celle basée la sur maximisation d'un indice (par exemple l'indice de Rand) ; celle par vote majoritaire, et celle basée sur les matrices association des partitions. C'est à cette dernière qu'on s'intéresse ici.

Dans cette approche, on considère un ensemble de T partitions $P = \{P_1, P_1, P_2, ..., P_T\}$ différentes d'un même ensemble de M observations. Ces partitions sont les résultats des multiples partitionnements pouvant provenir de plusieurs applications (initialisations) d'un même algorithme de classification, différents algorithmes sur le même jeu de données (notre cas) ou d'un même algorithme sur différents ensembles de variables décrivant les M individus.

Pour chaque partition P_t, on définit un tableau disjonctif contenant les indicatrices $H(P_t)$ des classes de la partition et une matrice d'association ou d'adjacence $M(P_t)$. La matrice d'association est une matrice $(M \times M)$ qui contient 1 si les deux individus i et j se trouvent dans la même classe, 0 sinon.

La matrice de connectivité $M(P_t)$ est obtenue par $M(P_t) = H(P_t)H(P_t)^{'}$ où $H(P_t)^{'}$ désigne la transposée de $H(P_t)$. On remarque qu'il s'agit d'une matrice symétrique positive contenant que des 1 en diagonale. Le nombre K de classes peut être différent d'une partition à une autre.

On définit la matrice d'association $\tilde{M}$ qui est une simple moyenne des matrices d'association par :

$$\tilde{M_{ij}} = \sum_{t=1}^{T} w_t M_{ij}(P_t) \tag{1}$$

Elle représente l'association moyenne entre deux observations (i, j). Dans une approche consensus simple (NMF), les poids sont donnés par $w_t = \frac{1}{T}$ pour toutes les partitions contributives. Par contre, l'approche pondérée repose sur la détermination et la recherche des poids w_t, en fonction de la qualité et la particularité de chaque partition. La matrice d'association s'exprime donc comme une moyenne pondérée des matrices d'association en fonction de ces poids. C'est le cas de "Weighted Nonnegative Matrix Factorization (WNMF) proposé par Ding et al. [4].

Cette approche est pertinente dans la mesure où certaines partitions peuvent paraître particulières par rapport aux autres. Dans ce cas, accorder le même poids à toutes les partitions peut biaiser la partition compromis obtenue.

Dans le cadre de ce travail, nous allons nous intéresser à l'approche WNMF. Cette approche permet la recherche simultanée de la partition compromis et les poids associés aux partitions contributives.

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

2.1.1 Les données

Cette étude est basée sur les articles publiés sur les quatre plateformes de bases 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, 10 237 publications ont été considérées pour l'étude.

2.1.2 Pré-traitement

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

- la suppression des 'stopword' qui consiste à supprimer tous les articles, pronoms et autres mots qui n'ont pas de sens pour notre analyse ;
- la tokenization 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 regrouper tous les mots en leur forme de base. (par exemple transformer tous les verbes conjugués en forme infinitif) ;
- 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[8]) 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 sparse 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 des paramètres et du corpus

Dans cette analyse, nous nous sommes basé sur le résumé de chaque article. Par ailleurs, puisqu'il s'agit d'une analyse non supervisée, et que nous n'avons pas l'information sur le nombre de sujets *a priori*, nous avons choisi un K optimal pour chaque modèle d'analyse au regard d'une mesure de cohérence CV [10] des sujets extraits.

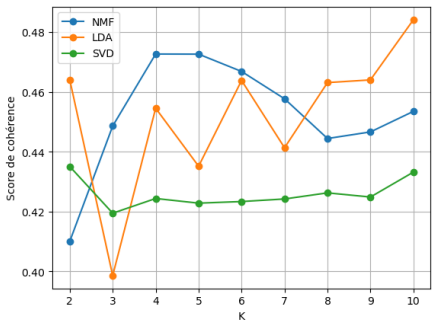

FIGURE 2 – Variation du score de cohérence en fonction du nombre de sujets K pour chaque modèle d'analyse

L'analyse de ce score en fonction du nombre de sujets K (FIGURE 2) pour chaque modèle nous permet de retenir les valeurs de K suivantes (4, 10, 2) respectivement pour NMF, LDA et k-SVD.

2.2 Résultats

Dans cette section, nous ferons, dans un premier temps, une analyse des sujets obtenus. Ensuite, nous évaluerons la qualité des partitions obtenues en nous basant sur la matrice θ. Une comparaison entre ces partitions sera faite en terme d'indices de Rand.

Une dernière analyse portera sur l'évaluation quantitative de la partition consensus obtenues à partir de l'approche WNMF.

2.2.1 Analyse des sujets

Une première analyse des sujets obtenus basée sur les TABLES 3, 1 et 5 permet de voir qu'on arrive à extraire des sujets cohérents traitant les thèmes biais, équité et explicabilité des algorithmes de ML. Par exemple, on peut voir que le sujet 3 extrait par la LDA concerne l'équité algorithmique dans un contexte de prise de décision où le besoin d'explicabilité et de compréhension de cette décision par l'humain se pose. Un autre constat est le fait que certains sujets sont extraits à la fois par les trois modèles d'analyse.

La TABLE 2 montre les résultats sur l'analyse quantitative de la qualité des sujets extraits au moyen des mesures de cohérences classiques. On note que la LDA a souvent une meilleure qualité. Par exemple pour l'indice UMASS, plus celui est faible, meilleure est la qualité des sujets obtenus en terme cohérence. Sur cet indice, on note que la LDA a la plus faible valeur. De même, lorsqu'on regarde les valeurs de CV, on note que l'approche LDA a une plus grande valeur (0.48). Ce qui signifie que ses résultats ont une meilleure qualité en termes de cohérence au regard de cet indice.

Sujet1	Sujet2
bias	bias
feature	attentional
fairness	cognitive
propose	negative
performance	participant
prediction	exchange
image	stimulus
classification	magnetic
analysis	find
base	positive

TABLE 1 – Description des sujets par les 10 mots les plus significatifs. Il s'agit des résultats de l'approche k-SVD.

		Indices de cohérence			
		UMASS	CV	UCI	NPMI
Modèles	NMF	-10.70	0.47	-7.16	-0.26
	LDA	-11.14	0.48	-7.51	-0.27
	k-SVD	-8.21	0.44	-6.09	-0.22

TABLE 2 – Qualité des sujets extraits selon les indices de cohérence classiques.

Sujet1	Sujet2	Sujet3	Sujet4
feature	bias	explanation	fairness
image	attentional	user	fair
propose	cognitive	explainable	algorithmic
classification	negative	decision	group
performance	find	explainability	metric
accuracy	patient	human	bias
problem	participant	prediction	problem
neural	attention	research	discrimination
prediction	risk	explain	framework
base	positive	trust	privacy

TABLE 3 – Description des sujets par les 10 mots les plus significatifs pour l'approche NMF

2.2.2 Comparaison des 3 partitions

Dans cette analyse, on s'est intéressé aux partitions de documents fournies par les trois approches d'analyse. L'objectif est de voir si on arrive à retrouver des structures similaires de partitions des trois approches. Pour analyser cette similarité, nous avons utilisé l'indice de Rand ajusté (ARI) [7] qui permet de quantifier la similarité entre deux partitions d'une même population.

L'analyse de la matrice des ARI fournit par la FIGURE 3 permet de constater une similarité entre les partitions de NMF et de la LDA. Cette similarité est plus faible pour la partition de la k-SVD.

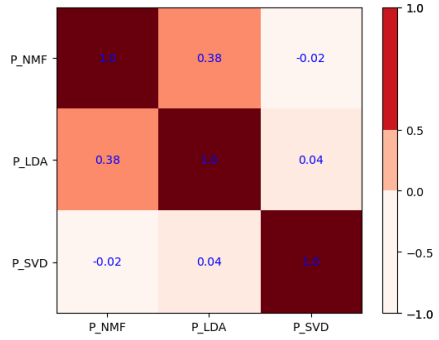

FIGURE 3 – Similarité des partitions au sens de l'indice de rand ajusté.

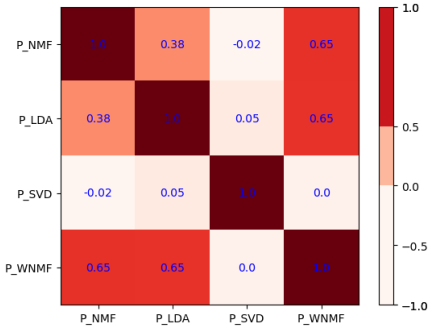

FIGURE 4 – Similarité entre les partitions contributives et la partition consensus au sens de l'indice de rand ajusté.

2.3 Résultats du consensus de partitions

Dans cette section, nous allons décrire les résultats fournis par le modèle de consensus de partitions en les comparant avec les partitions contributives obtenues grâce aux trois modèles de l'analyse initiale. Puisq'on considère l'approche LDA comme modèle de base, le nombre de clusters $K = 8$ sera considéré. Tout d'abord on constate que dans le consensus pondéré, le poids accordé à la partition de k-SVD est très faible en raison de sa forte différence avec les deux autres comme déjà soulignée TABLE 4. Ainsi, cette partition aura une plus faible contribution dans le processus de consensus.

Partitions	NMF	LDA	k-SVD
Poids	0.50	0.49	0.01

TABLE 4 – Poids accordé à chaque partition dans le processus de consensus WNMF. Ces poids sont fournis par l'algorithme de WNMF.

Une analyse de la similarité entre la partition consensus et les partitions contributives permet de noter une plus forte ressemblance entre la partition consensus et les partitions fournies par NMF et LDA. Ceci peut être expliqué par la différence de poids accordés aux différentes partitions. En effet, on a noté que ce poids est très faible pour la partition obtenue à partir de la k-SVD.

Conclusion et perspectives

Le travail proposé illustre l'intérêt des approches de traitement de langage naturel pour synthétiser, résumer, et même organiser une bibliographie 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 pratique, on est souvent emmené à faire un choix entre les modèles existants. Dans cette situation, nous proposons de faire recours à une approche consensus des résultats des modèles permettant ainsi d'éviter ce dilemme.

Cependant, cette approche permet uniquement de partitionner les documents et non les mots décrivant chaque classe comme le fait la LDA par exemple. Ainsi, dans nos futurs travaux, il serait intéressant de proposer une approche de consensus qui fournit également une relation entre chaque classe de documents et l'ensemble des mots. Ceci pourrait permettre d'interpréter facilement chaque classe de documents.

Références

[1] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai) : Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58 :82–115, 2020.

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

[3] Gosse Bouma. Normalized (pointwise) mutual information in collocation extraction. In *Proceedings of the Biennial GSCL Conference*, pages 31–40, 2009.

[4] Chris H. Q. Ding, Tao Li, and Wei Peng. On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. *Comput. Stat. Data Anal.*, 52(8) :3913–3927, 2008.

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

[6] N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness : Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2) :217–288, 2011.

[7] L. Hubert and P. Arabie. Comparing partitions. *Journal of classification*, 2(1) :193–218, 1985.

[8] 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.

[9] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755) :788–791, 1999.

[10] 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.

[11] 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.

[12] Alexander Strehl and Joydeep Ghosh. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.*, 3 :583–617, 2002.

A Description des 10 sujets extraits par l'approche LDA

sujet1	**sujet2**	**sujet3**	**sujet4**
bias	cell	explanation	fault
magnetic	receptor	fairness	history
exchange	signal	explainability	contain
substrate	protein	decision	threat
field	agonist	explainable	particle
device	drug	user	stress
measurement	activation	research	dependence
film	bind	human	science
phase	distinct	algorithmic	resistance
property	efficacy	technology	coverage

sujet5	sujet6	sujet7
patient	attack	bias
healthcare	security	cognitive
clinical	threat	participant
medical	adversarial	attentional
disease	judgment	gender
diagnosis	lime	risk
health	sensor	group
cancer	dnn	patient
care	behaviour	individual
treatment	resistance	find

sujet8	sujet9	sujet10
feature	recommendation	privacy
propose	item	fairness
classification	sentiment	traffic
image	recommender	federate
performance	user	resource
prediction	news	protocol
bias	social_medium	user
accuracy	medium	scheduling
problem	political	communication
neural	rating	throughput

TABLE 5 – Description des sujets par les 10 mots les plus significatifs pour l'approche LDA.