CESI LINEACT le cnam

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[1], G. Saporta [1]
{mlndao, gyouness}@cesi.fr
{ndeye.niang_keita, gilbert.saporta}@cnam.fr

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

July 4, 2023

CESI LINEACT

Introduction

Contexte

- Une croissance exponentielle de la littérature autour du biais, de l'équité et de l'explicabilité en IA à partir de 2017.
- Diversité de sujets traités par cette littérature
- Divergence dans certaines notions telle que la définition et la quantification de l'équité en Machine Learning.

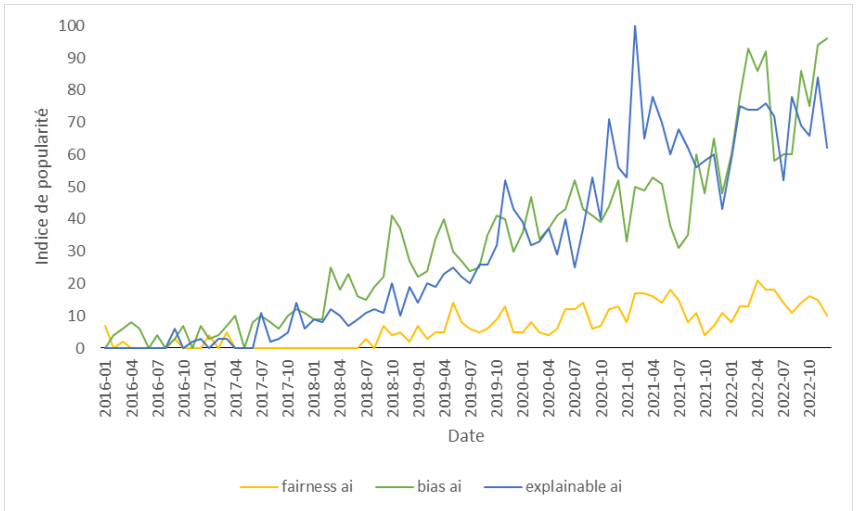


Figure: Popularité des termes ex- plainable XAI , Bias XAI et Fairness XAI dans le monde depuis 2016 selon Google Trends

Introduction

Objectifs

Proposer une analyse systématique (Topic Modeling) d'un échantillon de cette littérature en se basant principalement sur le modèle LDA (Blei, Ng, and Jordan, 2003) pour :

- Extraire et analyser les différents sujets traités par cette littérature.
- Analyser l'évolution de la popularité de ces sujets depuis 2015.
- Analyser les clusters de documents selon les sujets extraits.

Méthodologie : Latent Dirichlet Allocation (LDA)

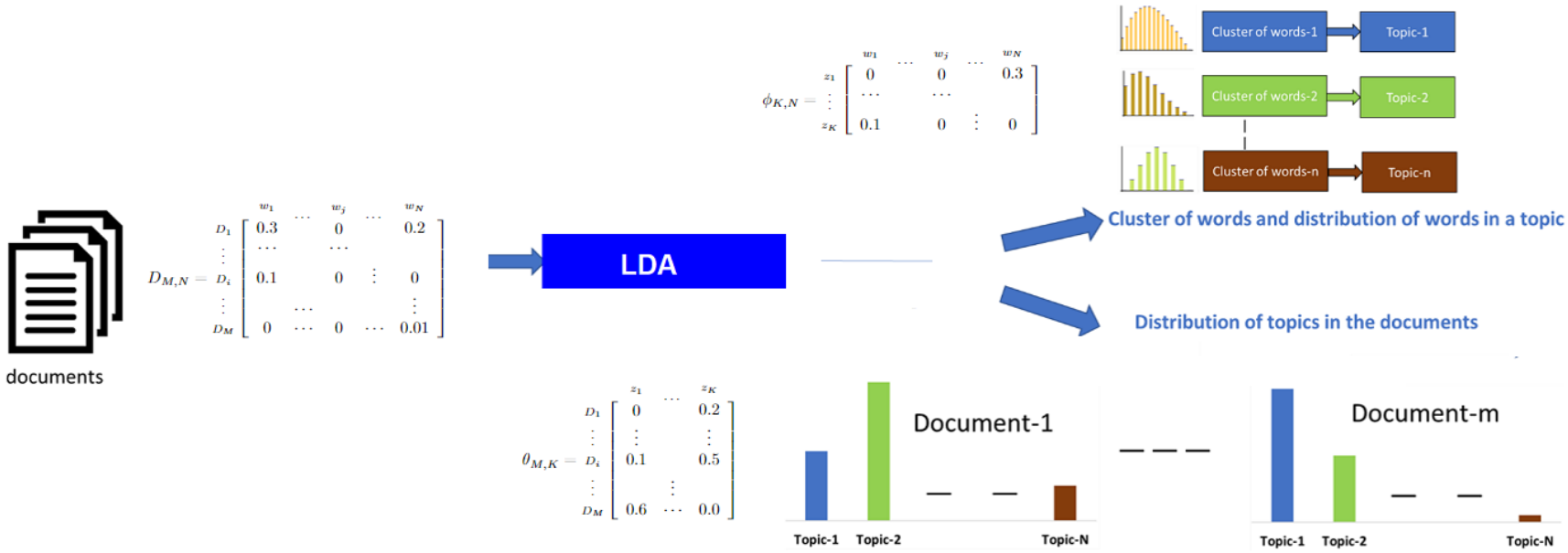


Figure: Principe de la LDA

Modèle probabiliste génératif permettant de trouver **la structure sous-jacente** d'un corpus.
Partant d'un corpus D (collection de documents), il estime :

- Matrice Φ la distribution des mots pour chaque sujet.
- La matrice Θ la distribution des sujets pour chaque document. .

Données

- Plateformes: arXiv, Springer, ScienceDirect et IEEE-Explorer.
- Articles publiés entre janvier 2015, décembre 2022.
- Concepts recherchés :
  - bias AND (machine learning OR data)
  - XAI AND (machine learning OR data)
  - fairness AND (machine learning OR data)
- 31 860 publications ont été considérées.

Processus d'analyse

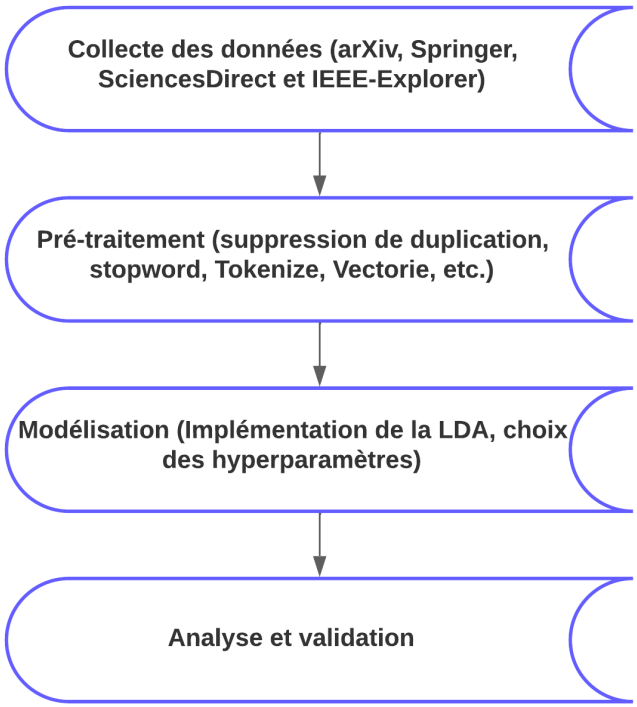


Figure: Processus d'analyse

- Métriques d'évaluation : Degré de similitude sémantique entre les mots les mieux notés dans un sujet.
  - *UMASS* : Unnormalized Measures of Association Strength (Mimno et al., 2011)
  - *C_V* : Coherence Value (Röder, Both, and Hinneburg, 2015).
- Validation : Comparaison de nos résultats avec BERTopic (Grootendorst, 2022) .

Choix du corpus et des paramètres

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: Résultats de l'analyse des différents corpus.

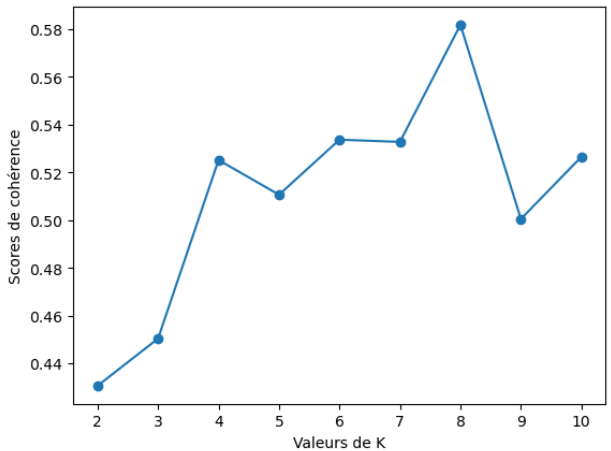


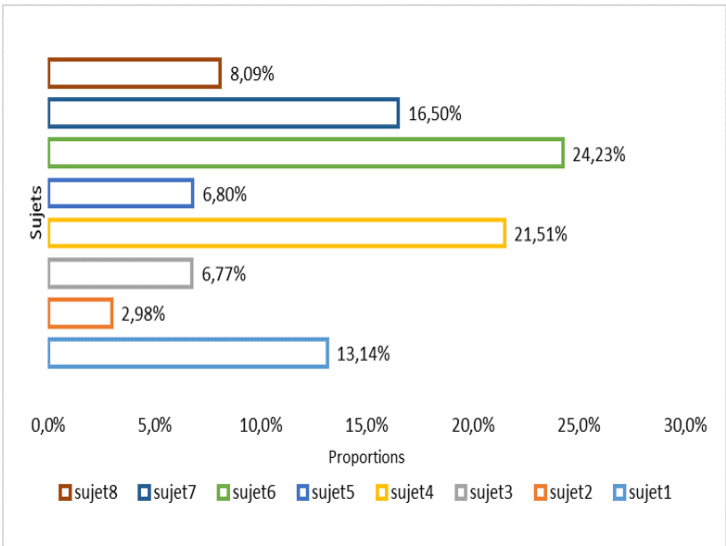
Figure: Variation du score de cohérence (C_V) en fonction du nombre de sujets K lorsque $\alpha = 0.91$ et $\beta = 0.91$ pour le corpus abstract et keywords.

- Corpus : résumé + mots-clés
- Nombre de sujets $K = 8$.

Résultats : Les 8 sujets extraits et leur popularité

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: Description des sujets par les 10 mots les plus significatifs.



Nom des sujets

1 : biais en science cognitive ; 2 : l'éthique et la confidentialité des données ; 3 : biais en génétique ; 4 : "boîtes noires" et données complexes ; 5 : l'équité en Cloud ; 6 : équité et explicabilité en ML ; 7 biais en télédétection ; 8 : biais en santé .

Figure: Répartition des publications entre les sujets en %

CESI LINEACT

Résultats : Évolution de la popularité des sujets depuis 2015

Forte croissance:

- **Equité et explicabilité en ML**
- **"Boîtes noires" et données complexes**

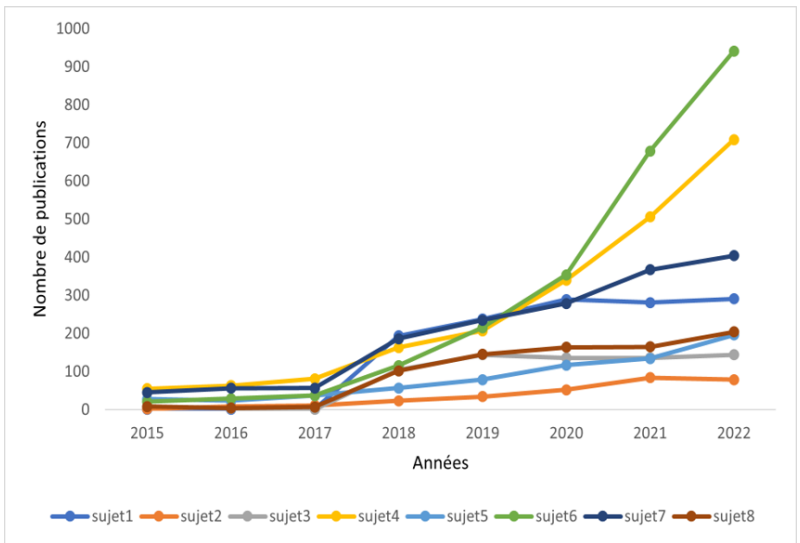


Figure: Évolution du nombre de publications par sujet depuis 2015

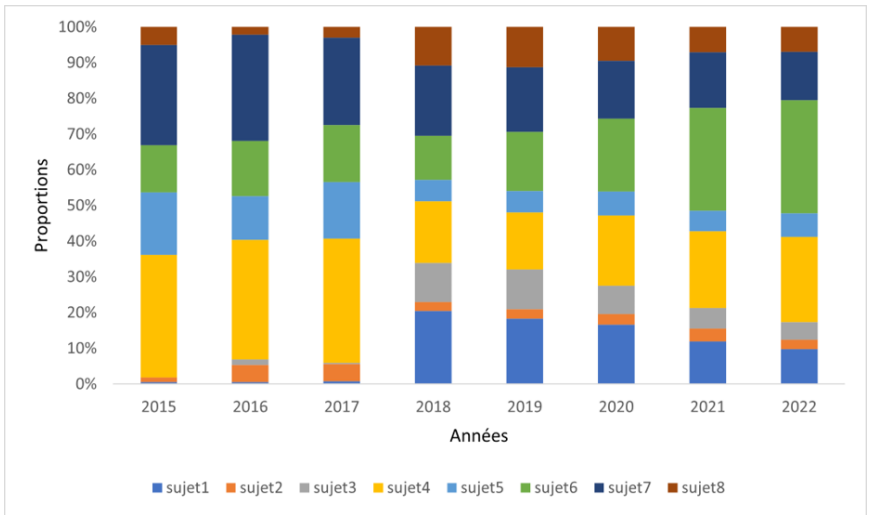


Figure: Évolution de la part des publications traitant chaque sujet depuis 2015

Résultats : Analyse des clusters de publications selon les sujets

- Une bonne séparation des clusters trouvés.
- Non popularité du sujet 2.
- Une proximité des sujets 4, 7 et 5.

Figure: Visualisation de la distribution des clusters par l'outil pyLDAviz, (Sievert and Shirley, 2014)

CESI LINEACT

Résultats : Description des clusters 4 et 6

Cluster 4 : données complexes et boîtes noires

- Les publications ayant les plus fortes probabilités d'appartenance au cluster 4 parlent des données complexes telles que les images et les boîtes noires comme les CNN .
- Certaines publications ont parlé du biais selon :
  - le type de données (images, graphe, audio, tabulaire);
  - selon le type de modèle d'analyse utilisé (supervisé, semi supervisé, non supervisé).

Cluster 6 : équité et explicabilité en ML

- Les publications ayant les plus fortes probabilités d'appartenance à ce cluster parlent de l'équité et de l'explicabilité en ML .
- Ces publications montrent un lien fort entre l'équité et l'explicabilité :
  - aspect humain ;
  - deux termes traités dans un contexte de prise de décisions sur la base des ML.

LINEACT

Résultats : Comparaison entre nos résultats et ceux de BERTopic

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

Table: Scores de cohérence des modèles BERTopic (Grootendorst, 2022) et LDA

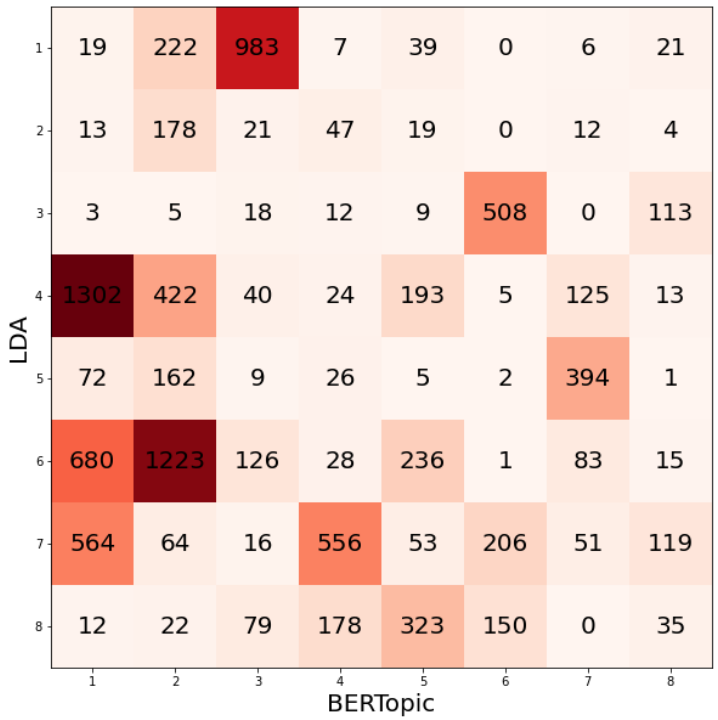


Figure: "Matrice de confusion" entre les clusters de LDA et BERTopic : correspondance quantitative entre les clusters obtenus selon LDA et BERTopic.

CESI LINEACT

Conclusion

Ce travail a permis de :

- Identifier 8 sujets associés chacun à un cluster de publications autour du biais de l'équité et de l'explicabilité.
- Réorganiser de cette bibliographie pour une meilleure exploitation
- Montrer la pertinence des approches de NLP dans l'exploitation des données d'archives dans un contexte de données massives.

Perspectives

- Approfondir la comparaison entre LDA et BERTopic.
- Utiliser les clusters de documents obtenus pour organiser davantage notre étude bibliographique.

References I

- Blei, David M, Andrew Y Ng, and Michael I Jordan (2003). "Latent Dirichlet Allocation". In: *Journal of machine Learning research* 3.Jan, pp. 993–1022.
- Grootendorst, Maarten (2022). "BERTopic: Neural topic modeling with a class-based TF-IDF procedure". In: *arXiv preprint arXiv:2203.05794*.
- Mimno, David et al. (2011). "Optimizing semantic coherence in topic models". In: *Proceedings of the 2011 conference on empirical methods in natural language processing*, pp. 262–272.
- Röder, Michael, Andreas Both, and Alexander Hinneburg (2015). "Exploring the space of topic coherence measures". In: *Proceedings of the eighth ACM international conference on Web search and data mining*, pp. 399–408.

References II


Sievert, Carson and Kenneth Shirley (2014). "LDAvis: A method for visualizing and interpreting topics". In: *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, pp. 63–70.

Merci pour votre attention