Vieux mais pas obsolètes : Bag-of-Words vs. Embeddings en modélisation thématique

Jean-Charles Lamirel[1,2], Francis Lareau[3], Christophe Malaterre[4], Thibault Prouteau[2]

[1]Université de Strasbourg, [2]MosAIk-LORIA, [3] Université de Sherbrooke / UQAM, [4] UQAM, CIRST

lamirel@loria.fr

Résumé

Cet article compare quatre méthodes de modélisation thématique — LDA, CFMf, Top2Vec et BERTopic — sur un corpus de 16 917 articles de philosophie des sciences, selon des critères de cohérence, de diversité et de rappel. Les résultats montrent que Top2Vec excelle en cohérence et diversité mais échoue en rappel et interprétabilité. BERTopic surpasse légèrement LDA en cohérence mais pas en rappel. CFMf, grâce à une adaptation métrique angulaire du clustering GNG, offre le meilleur équilibre entre toutes les dimensions. Ces résultats démontrent la compétitivité durable des approches BOW et soulignent la nature modulaire des pipelines de modélisation thématique.

Mots-clés

Modélisation thématique, Bag-of-Words, embeddings, LDA, CFMf, BERTopic, Top2Vec, cohérence, diversité, rappel.

Abstract

This paper compares four topic modeling approaches — LDA, CFMf, Top2Vec, and BERTopic — on a corpus of 16,917 philosophy of science articles across coherence, diversity, and recall metrics. Results show Top2Vec leads on coherence and diversity but lags on recall and interpretability; BERTopic marginally outperforms LDA on coherence; CFMf, using an angular GNG adaptation, achieves the best overall balance. Statistical BOW-based models remain competitive against modern embedding methods.

Keywords

Topic modeling, Bag-of-Words, embeddings, LDA, CFMf, BERTopic, Top2Vec, coherence, diversity, recall.

1 Introduction

La modélisation thématique est un outil important de l'analyse computationnelle de corpus textuels. Les approches classiques fondées sur la représentation Bag-of-Words (BOW) — notamment LDA [2] — des documents textuels ont longtemps dominé le domaine. L'émergence de modèles basés sur les plongements lexicaux/embeddings (Top2Vec [1], BERTopic [5]) a relancé la question de leur pertinence.

Cet article prolonge des travaux antérieurs sur CFMf [7], une méthode BOW combinant la maximisation des traits (Feature Maximization) et le clustering neuronal à base de gaz croissants (GNG). L'étude poursuit trois objectifs : (i) corriger les défauts résiduels de CFMf, (ii) étendre la comparaison à BERTopic, (iii) étudier la modularité des pipelines de modélisation thématique.

2 Méthodes comparées

LDA [2] est un modèle génératif statistique représentant chaque document comme un mélange de thèmes, chacun défini par une distribution sur les mots. Il utilise une représentation initiale BOW des documents textuels et une inférence bayésienne (échantillonnage de Gibbs ou inférence variationnelle).

CFMf [7] utilise aussi une représentation BOW des documents et combine la maximisation des traits (F-mesures de traits) (6] pour le ranking des mots avec le clustering GNG [4] pour le regroupement. GNG est un algorithme winner-take-most adaptatif qui apprend la topologie des données sans fixer le nombre de clusters a priori. Une version angulaire de GNG est introduite dans ce travail : à chaque étape d'apprentissage, les vecteurs prototypes des clusters gagnants sont renormalisés après la mise à jour des poids. Cette renormalisation projette les prototypes sur la sphère unité, rendant la compétition entre neurones (prototypes) insensible au nombre de mots présents dans les documents et évitant ainsi les clusters attracteurs [7].

Top2Vec [1] utilise Doc2Vec pour créer des plongements sémantiques denses de mots et documents. HDBSCAN fait émerger des clusters denses sans fixer K a priori ; les centroïdes de clusters servent de représentants thématiques et le ranking des mots repose sur leur proximité cosinus au centroïde.

BERTopic [5] exploite des plongements contextuels profonds (ici réalisés avec le modèle Transformer stella_en_1.5B_v5), HDBSCAN pour le clustering et la réassignation des outliers, et une pondération c-TFIDF pour le ranking des mots. BERTopic illustre ainsi la modularité des pipelines en combinant des plongements Transformer (pour le clustering) et une représentation BOW (pour le ranking des top-mots).

3 Protocole expérimental

Le corpus comprend 16 917 articles en texte intégral issus de huit revues de référence en philosophie des sciences (1930–2017) [8]. Après tokenisation, étiquetage morphosyntaxique et lemmatisation (TreeTagger), seuls les noms, verbes, adverbes et adjectifs apparaissant dans au moins 50 phrases sont conservés. Les modèles sont évalués pour $K \in [5, 100]$ selon quatre métriques complémentaires :

– Cohérence C_V [9] : co-occurrence des mots d'un thème ;

– Diversité thématique : ratio de mots-clés uniques sur le total des mots-clés ;

– Rappel interne micro (mIR) : capacité moyenne des mots-clés à rappeler les documents de leur thème ;

– Rappel interne joint micro (mJIR) : couverture conjointe de l'ensemble du corpus par tous les mots-clés.

Les thèmes LDA sont alignés sur les clusters de CFMf, Top2Vec et BERTopic en affectant leurs documents de manière stricte (selon leur thème dominant), ceci pour faciliter la comparaison qualitative.

4 Résultats

La **cohérence** croît avec K pour tous les modèles, atteignant un plateau vers K=20–30 pour LDA, CFMf et Top2Vec, et plus tardivement pour BERTopic. Top2Vec atteint les valeurs les plus élevées ($C_V \approx 0{,}80$ dès K=20), suivi de CFMf ($C_V > 0{,}70$). LDA affiche les scores les plus bas (plateau ≈

0,55), légèrement dépassé par BERTopic aux valeurs élevées de K.

La **diversité thématique** décroît à mesure que K augmente. Top2Vec maintient la diversité la plus haute (> 0,95 pour tout K). CFMf se classe deuxième (0,80–0,90). LDA et BERTopic convergent vers 0,65 à partir de K=30.

Le **rappel** révèle un tableau radicalement différent. Top2Vec enregistre les scores mIR les plus faibles (< 0,35), révélant une déconnexion entre ses mots-clés et ses documents. LDA domine (mIR > 0,80), suivi de BERTopic (0,70–0,80) et de CFMf (≈ 0,60 pour K > 25). Il faut noter que la valeur mIR présente une signification particulière pour détecter les modèles dont la couverture documentaire est faible, comme c'est le cas pour Top2Vec, lorsqu'elle est nettement basse. En revanche, l'interprétation de valeurs mIR élevées comme indicateurs de performances supérieures doit être abordée avec prudence. De telles valeurs élevées peuvent résulter de la présence de termes fréquents mais sémantiquement peu informatifs dans les top-mots des topics, ce qui nuit à l'interprétation de ces derniers.

Pour le mJIR, LDA, BERTopic et CFMf atteignent tous ≈ 1, indiquant une couverture quasi-totale du corpus. Top2Vec plafonne à 0,90, laissant ≈ 10 % des documents non rappelés quelle que soit la valeur de K.

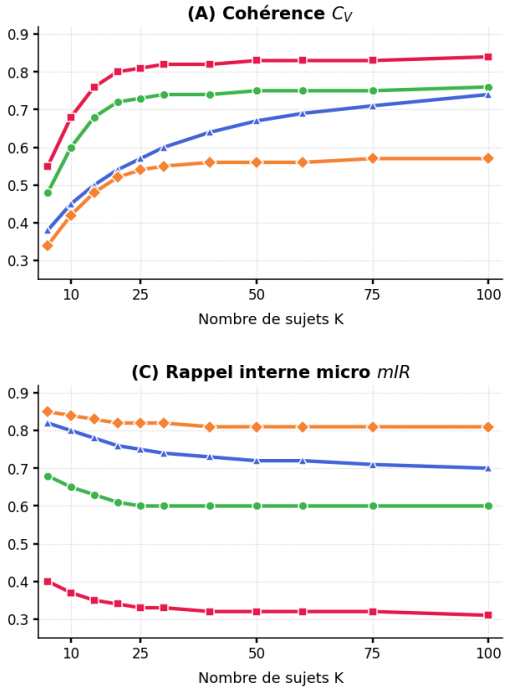

Fig. 1. Comparaison des modèles en fonction de K : (A) cohérence CV, (C) rappel interne micro mIR. Modèles : Top2Vec (carrés rouges), CFMf (cercles verts), BERTopic (triangles bleus), LDA (losanges oranges).

4.3 Interprétabilité qualitative

L'alignement thématique (K=25) montre une bonne convergence descriptive globale. Néanmoins, les top-mots thématiques de Top2Vec incluent fréquemment des noms d'auteurs et des termes très spécifiques, réduisant leur interprétabilité. CFMf génère des descripteurs précis et bien délimités — distinguant par exemple relativité et mécanique quantique — avec quelques noms d'auteurs résiduels. BERTopic produit un mélange de descripteurs clairs et de descripteurs génériques. LDA tend à produire un nombre important de descripteurs génériques

5 Discussion

Les résultats mettent en évidence des compromis fondamentaux entre approches. Les modèles à plongements (Top2Vec, BERTopic) tirent parti de représentations sémantiques riches mais présentent des faiblesses importantes en rappel pour Top2Vec, en cohérence pour BERTopic, et en interprétabilité pour les deux. Les approches BOW, en particulier CFMf, offrent un meilleur compromis vis-à-vis de la couverture, de la cohérence et de l'interprétabilité. Un aspect clé concerne la modularité des pipelines : les composantes de prétraitement, vectorisation, clustering et ranking de mots sont en grande partie indépendantes et combinables. L'adaptation angulaire de CFMf illustre comment une modification ciblée d'un seul composant peut éliminer un défaut structurel sans modifier le reste du pipeline. Des travaux futurs exploreront des stratégies de réassignation des données marginales plus fines, susceptibles d'améliorer davantage les petites classes, ainsi que la prise en charge d'embeddings par le clustering GNG dans la chaine CFMf. Parmi les limitations : le corpus est restreint à un seul domaine disciplinaire ; de nombreuses autres approches restent à explorer. Ce travail met également en évidence les limites des métriques automatiques pour mesurer la qualité des thèmes, y compris les plus couramment utilisées, comme la cohérence |9]. Cela suggère l'intérêt de les compléter par une évaluation humaine et une évaluation fondée sur les LLM, dans le cadre d'un protocole d'évaluation mixte et parallèle.

6 Conclusion

Aucune approche ne domine uniformément sur toutes les dimensions évaluées. Top2Vec excelle en cohérence et diversité mais échoue en rappel et interprétabilité. LDA et BERTopic privilégient la couverture documentaire, mais peinent en interprétabilité et en cohérence. CFMf, grâce à son adaptation angulaire, réalise cependant le meilleur compromis global et génère des mots-clés hautement interprétables. Ces résultats réhabilitent les modèles BOW face aux méthodes à plongements et soulignent le potentiel d'approches hybrides exploitant la modularité des pipelines — une piste prometteuse pour des travaux futurs.

Références

[1] D. Angelov. Top2Vec: Distributed Representations of Topics. arXiv:2008.09470, 2020.

[2] D. M. Blei, A. Y. Ng, M. I. Jordan. Latent Dirichlet Allocation. JMLR, 3:993–1022, 2003.

[3] J. Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers. Proc. NAACL-HLT, pp. 4171–4186, 2019.

[4] B. Fritzke. A growing neural gas network learns topologies. NIPS, 7, 1994.

[5] M. Grootendorst. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv:2203.05794, 2022.

[6] J.-C. Lamirel, N. Dugué, P. Cuxac. New efficient clustering quality indexes. IJCNN, pp. 3649–3657, 2016.

[7] J.-C. Lamirel, F. Lareau, C. Malaterre. CFMf topic-model: Comparison with LDA and Top2Vec. Scientometrics, 129:6387–6405, 2024.

[8] C. Malaterre, F. Lareau. The early days of contemporary philosophy of science. Synthese, 200(3):242, 2022.

[9] M. Röder, A. Both, A. Hinneburg. Exploring the Space of Topic Coherence Measures. WSDM, pp. 399–408, 2015.